A New Approach for Extracting Inter-word Semantic Relationship from a Contemporary Chinese Thesaurus

by

Lam Sze-sing

Systems Engineering and Engineering Management

Department

Submitted to the Chinese University of Hong Kong in partial fulfillment of requirements for the degree of

Master of Philosophy

June, 1995

Abstract

In this thesis, I propose a model to extract the knowledge embedded in a contemporary Chinese thesaurus 《同義詞詞林》, CILIN. This extraction process is crucial to the application of thesaurus to natural language processing (NLP). The model directly computes the *connection weights* between the *semantic classes* in the thesaurus by measuring *weak* as well as the *strong relationships* between them. The connection weights are then used to estimate the *semantic relationship* between two *words*. I have shown in this study that the traditional *Simple Co-occurrence approach* can only measure some *semantic relationship* between *semantic classes* in the thesaurus. The approach is too restrictive for it measures only the *strong relationship* between semantic classes and fails to account for their *weak relationship*. Consequently, many null relationship is produced.

The model is tested using the Noun-Verb-Noun (N-V-N) compound word detection problem. This problem is currently widely studied for simplifying the syntactic analysis process of Chinese sentences. Testing results indicate that the new model can improve *recall* from 2%, based on the simple co-occurrence approach, to 55%.

The model is finally used to tackle the formidable task of word-sense disambiguation. A linguistic based word-sense disambiguation algorithm, LSD-C,

is proposed for resolving the meanings of ambiguous words using both the context of the document which they are extracted from, and the meaning of the words as defined in a standard Chinese dictionary, namely《現代漢語詞典》. Application of the algorithm to a set of Chinese articles selected from a collection of local Chinese newspapers shows that it offers an accuracy rate comparable to existing English word-sense disambiguation algorithms. Compared to the existing English systems, the proposed algorithm is more comprehensive: (1) the testing samples are domain non-specific, (2) the ambiguous words are not pre-selected, and (3) the disambiguation is on the level of *polysemy* (i.e. fine-grained differences in meaning). Moreover, it is simpler to implement for it employs only the semantic knowledge provided by the thesaurus.

Acknowledgements

I would like to thank all of the people who have helped and guided me through my graduate work.

I would particularly like to express my deepest gratitude to my advisors, Professor Vincent Lum and Dr. K.F. Wong, for their support, their helpful suggestions, and willingness to help me out whenever I wandered in the research path. Without their advise, I would not be able to accomplish this task. Their valuable comments on every details, particularly on the formulation of the theme, have helped make this thesis possible.

My final and greatest thanks go to my family. Without their support and bearing, I would have surrendered my study in a number of occasions when I felt miserable and helpless in my research. When I needed them, they have always been there, especially my wife who not only allowed me to work over-night without the slightest complaint but also tagged some of the testing data for me.

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

CHAPTER 1 INTRODUCTION

1.1 Introduction

Collins Cobuild Dictionary [Cobuild87] defines "knowledge" as *information and understanding about a subject which a person has in his or her mind or which is shared by all human beings.* It is obvious from this definition that knowledge is fundamental to human communication. For daily communication, humans use a common natural language. Learning a natural language is therefore rudimentary. Fundamental in the learning process, knowledge of words and of their associated meaning (i.e. *semantic*) are most important, especially in written communication [Fromkin93]. Without this information, it is impossible to interpret and understand correctly the materials written by others. Therefore, to construct a computer model to process and understand natural language text - this process commonly referred to as Natural Language Processing (NLP), one has to solve the problem of where and how to acquire the knowledge of words.

In the past, many of the well known NLP systems such as LUNAR, LIFER, SHRDLU [Patterson90] and FRUMP [DeJong89] made extensive use of *hand-crafted knowledge bases*[1]. As most of the required knowledge was hand-crafted, these systems were confined to a narrow domain. It was costly and impractical to

[1] The knowledge required for the running of the system is manually collected and encoded into a knowledge base.

expand these systems to a more general domain. In consequence, they suffered from the well-known *scale-up* problem typically found in NLP systems. To overcome the knowledge acquisition bottleneck, the CYC [Lenat86] project even once employed ten to twenty programmers to hand-craft a knowledge base from a selection of encyclopedia articles. This form of knowledge described the world and was complementary to the knowledge of word in NLP. Amsler [Amsler89] was concerned in how artificial intelligence (AI) techniques could be used to build a replacement of world knowledge required in performing the task of text understanding and argued that the bulk of this knowledge would have to be derived from dictionaries which had served us well for a few hundred years.

Recently, more and more NLP researches suggested methods for acquiring knowledge automatically [Boguraev89b, Wilks89, Harder91, Tsutsumi91, Antonio94, Jacobs94]. Computer word knowledge acquisition approaches can be broadly divided into two classes: *linguistic-based* and *corpus-based.* As their names suggest, the major source of knowledge in the corpus-based approach is the *corpus* (i.e. a collection of documents) and the same in the linguistic-based approach is the linguistic resources currently available to and designed for human, such as dictionary and thesaurus. In general, the corpus-based approach is more domain specific as its knowledge comes from documents written for a dedicated application domain. However, unlike dictionary and thesaurus, a corpus requires some form of preprocessing. This is often an expensive operation.

Wilks et al. [Wilks89] suggested that the dictionary provided not just knowledge about definitions of words but also knowledge about the world. They attempted to extract and use the semantic information from a machine readable dictionary (MRD) - the Longman Dictionary of Contemporary English (LDOCE). Similarly, Harder, and Tsutsumi [Harder91, Tsutsumi91] proposed different methods to use the knowledge in the LDOCE for disambiguating word senses. The linguistic resources, such as dictionary and thesaurus, are knowledge serves for dealing with many NLP problems. It is worthwhile to explore these knowledge sources in depth in view of their availability and generality.

However, both dictionary and thesaurus are tailored for human consumption and are unsuitable for NLP. Since they are collected and recorded by lexicographers for human readers to solve different language problems, they contain knowledge about the language and the world essential for different tasks in NLP. The approaches to acquire the knowledge from them largely depend on the type of knowledge required by the target applications.

Different from Western languages, such as English and French, Chinese is more semantically driven [Lua93b, Tang94]. This is the major reason why it is so difficult for researchers to put forward a complete set of formal grammatical rules in spite of years of research efforts. Formulation of a Chinese sentence is restricted more by the semantic relationship of the words rather than by their syntactical relationship. For example, amongst the words 聽覺, 熒光屏 and 視力 (hearing, screen, and vision, respectively), the word 視力 (vision) is semantically related to

the word 熒光屏 (screen) but not to the word 聽覺 (hearing). As a result, screen (熒光屏) can affect (影響) vision but hearing (視力) cannot. Using these semantic information for analysis of the following 2 sentences:

(a)	計算機 熒光屏 對 視力 也 有 不良 影響。 Computer screen also has adverse effect on vision.
(b)	聽覺 對 視力 也 有 不良 影響。 Hearing also has adverse effect on vision.

We can confidently determine that sentence (a) is correct and (b) is not even though both sentences are syntactically correct. Thus, knowledge about the semantic relationship of words is crucial to the understanding of natural language sentences and Chinese is no exception. This knowledge is applicable to many NLP issues such as *Word-Sense Disambiguation*, *Information Retrieval, Compound Word Detection*, etc.

In this study, I conduct a detailed study on how to extract knowledge about the *inter-word semantic relationship* (i.e. the semantic relationship of words) from a contemporary Chinese thesaurus. The thesaurus clusters the semantically related words in classes (i.e. *semantic classes*) and represents the hierarchical relationship of these semantic classes as a tree. However, this hierarchical relationship is insufficient for estimating the inter-word semantic relationship. Only association between the semantic classes can be estimated. Lua K.T. [Lua93b] has suggested the conventional *Simple Co-occurrence approach* for computing such an association. Unfortunately, simple co-occurrence statistics can only measure the direct link

between semantic classes through counting the number of words shared between them. In general, simple co-occurrence statistics can only represent the *strong relationship* (i.e. a direct link) between the semantic classes and not their *weak relationship* (i.e. an indirect link) [Agosti92]. Consequently, thesaurus information extracted using Lua's approach is partial and incomplete.

1.2 Statement of Thesis

This thesis presents a new approach to automatically extract the *inter-word semantic relationship* from a contemporary Chinese thesaurus《同義詞詞林》(CILIN) [Mei83]. This approach is an extension of the conventional Simple Co-occurrence approach. It measures the *weak relationship* as well as the *strong relationship* between the *semantic classes* in the thesaurus for estimating the *connection weight*. The *semantic relationship* between any pair of words is then derived from the connection weight found. Simple co-occurrence method can only measure the *strong relationship* and miss out the weak one. Effectively, not all the knowledge can be extracted from the thesaurus by simple co-occurrence. Therefore, the objective of this new approach is to overcome this shortcoming.

The testbed for the new approach is the Noun-Verb-Noun (N-V-N) compound word detection problem which is a NLP problem specific to Chinese. N-V-N word identification techniques are used to simplify the syntactic analysis in Chinese sentence. A new N-V-N compound word detection model is developed. It makes extensive use of the inter-word semantic relationship measure for evaluating the likelihood of the formation of a N-V-N compound from a given set of words. This

model is aimed at comparing the new approach with the simple co-occurrence method.

A word-sense disambiguation algorithm is also developed. It is used to illustrate how the inter-word semantic relationship can be applied to NLP. The algorithm resolves the word-sense ambiguity as defined in the standard dictionary using the semantic cue provided by the words at the sentence level.

The major contributions of these works are:

- ✧ Propose a fully automatic approach for extracting the inter-semantic relationship from a contemporary Chinese thesaurus. The new approach is superior to the simple co-occurrence model.

- ✧ Design and develop an algorithm to resolve the word-sense disambiguation problem in Chinese full-text using the inter-word semantic knowledge extracted from the thesaurus.

1.3 Organization of this Thesis

The organization of the remaining chapters of this thesis is as follows:

Chapter 2 - In chapter 2, related work in knowledge acquisition is discussed in details. Different approaches to extract application specific knowledge from various knowledge sources are summarized. The pro and cons of each of them are outlined. Current status of Chinese NLP is briefly described and the major reasons for studying the CILIN are stated.

Chapter 3 - In this chapter, the algorithm to acquire the inter-word semantic relationship from the CILIN and the design of the knowledge representation model to capture the semantic information obtained are given. Research efforts related to the design of the algorithm and the knowledge representation model are presented. The problems in the simple co-occurrence method for gathering semantic knowledge from the CILIN are elaborated.

Chapter 4 - This chapter outlines an experiment to compare the efficiency of the knowledge acquisition approach presented in this thesis with the simple co-occurrence method. The N-V-N compound word detection problem, a widely studied problem in Chinese NLP, is used in the experiment. The experimental results are evaluated and discussed.

Chapter 5 - A word-sense disambiguation model is proposed in this chapter to illustrate how to apply the inter-word semantic relationship derived from the CILIN to tackle a complicated problem in Chinese NLP. The stepwise procedure of the disambiguation algorithm is described and an example is given to demonstrate how the algorithm operates. The results of the example are analysed and discussed.

Chapter 6 - The last chapter summarizes this study and discusses about future researches.

CHAPTER 2 RELATED WORK

In this chapter, I will summarize some recent researches in the field of knowledge acquisition. In the first section, I will provide an overview on the importance of knowledge, the need for automatic knowledge acquisition techniques and the basic criteria to evaluate them. In the next two sections, I will attempt to categorized different knowledge acquisition techniques and then briefly describe their basic principles. Also, I will highlight their strengths and weaknesses.

2.1 Overview

By nature, language is the primary vehicle for people to communicate and to record information. Noam Chomsky in his book *Language and Mind* once said:

> *When we study human language, we are approaching what some might call the "human essence," the distinctive qualities of mind that are, so far as we know, unique to man ...* [Chomsky72]

Fromkin and Rodman [Fromkin93] even claimed that the possession of natural languages distinguishes humans from other animals. Today, the ability to understand and generate natural language is not only important to human but also to computers. In fact, we can trace back to as early as 1950s when people started to recognize the potential of Natural Language Processing (NLP) in computer.

The primary motivation in NLP is to develop a computer system for processing information expressed in natural language [Allen87, Cullingford86, Grishman86]. Major activities in these systems are to understand and, in some case, to generate natural language. In order to understand natural language, a NLP system must know a fair amount of knowledge about the language and the subject area. This kind of knowledge can be lexical, grammatical, semantical, and/or pragmatical. The major theoretical issues pertinent to the knowledge are how to acquire, to represent and to use it effectively on a computer. In the last two decades, virtually all NLP systems acquire knowledge by some hand-crafted approaches. These knowledge acquisition approaches are laborious. Due to time and cost limitations, such NLP systems are only furnished with knowledge of their intended application domains. This seriously limits the scope of usage of these systems.

Since they were designed with a specific purpose in mind, the application of these NLP systems to a new domain would require some additional knowledge. Because of this barrier, previous NLP systems could only handle toy cases. To tackle the knowledge acquisition bottleneck, researchers try to automate the process of knowledge acquisition. Many knowledge acquisition techniques have been devised for the purpose.

Wilks et al. [Wilks89] proposed to acquire knowledge directly by developing methods to transform the knowledge within dictionaries or encyclopedias into some format usable to NLP. Similarly, Bates et al. [Bates93] suggested a number of ways to acquire knowledge automatically:

- ✧ Using references such as on-line dictionary which one can use to find exact information.

- ✧ Using references such as an encyclopedia or a corpus of domain-relevant material, from which one can find or infer the information being sought. It may also mean using large volume of material as the source of probabilistic knowledge.

- ✧ Using heuristics and the information in the input itself such as the part of speech of the words surrounding an unknown word.

In general, these approaches can be characterized by the kind of references utilized. They can be broadly divided into *linguistic-based* and *corpus-based* approaches. The major source of knowledge in the linguistic-based approach is standard linguistic resources such as dictionaries and thesauri, and the major source of knowledge in the corpus-based approach, as its name suggests, is a corpus.

To extract knowledge from a piece of text, automatic knowledge acquisition must address the following three issues: *sufficiency*, *extricability*, and *bootstrapping* [Wilks93]. Sufficiency addresses the issue of whether the knowledge source is a strong enough knowledge base for the NLP task. Extricability concerns with whether it is possible to specify a set of computational procedures that can extract general and reliable semantic information from the knowledge source to a general format suitable for a range of subsequent NLP tasks. Bootstrapping refers to the process of collecting the initial information that is required by a set of computational

procedures for extracting semantic information. Bootstrapping methods can be *internal* or *external*. Internal methods obtain the initial information needed for their procedures from the knowledge source itself and the external methods obtain initial information from some other sources (e.g. pre-process a corpus by manually adding the part-of-speech tagger).

Generally, knowledge contents in a knowledge source discussed in the following sections are normally sufficient and extricable, at least to the respective application. Therefore, among the three factors, the sufficiency and the extricability are normally not the main concern if a knowledge source is selected appropriately. However, bootstrapping is an important issues in the design of a knowledge acquisition system. One reason to employ a knowledge acquisition model is to reduce the cost in acquiring the knowledge required by the NLP systems and the original idea to save time will be escalated if the bootstrapping process itself is a time consuming step. Therefore, the bootstrapping procedure of a knowledge acquisition approach is an indicator about the cost effectiveness.

Theoretically, both the dictionary and the corpus should contain knowledge about the language and about the world. They contain rich lexical information which extends well beyond simply defining part-of-speech and are capable of supporting a wide range of activities, both of theoretical interest and of practical importance [Boguraev89]. How to glean the knowledge and to represent it is largely a decision for the NLP problem on-hand. The knowledge so extracted is unique to the problem. Therefore, the overall design of a set of the knowledge

acquisition procedures are, in general, closely related to the NLP task under tackle and its corresponding knowledge representation model.

Comparing to English, Chinese NLP started much later. Its progress is relatively slow [Wang90, Zhang92, Yu93]. Most of the early effort in Chinese NLP was devoted to solving some of the fundamental issues that are unique to the Chinese language such as *word segmentation* and *syntactic categorization of* Chinese words. Only until recent years, the Chinese NLP community came up with standards in these issues [GB92, Tsinghua92]. Nowadays, similar to other languages, researchers in Chinese NLP have commenced to explore different areas in NLP including automatic knowledge acquisition from a corpus or from a standard dictionary. In order to have an overall view on what can be provided and how to obtain the knowledge from different kinds of knowledge sources, I will outline various most recent knowledge acquisition approaches in the next sections.

2.2 Corpus-Based Knowledge Acquisition

Recent researches in corpus-based knowledge acquisition cover a wide range of NLP specific knowledge including lexical, syntactic, semantic and even pragmatic knowledge. Weischedel et al. [Weischedel93] conducted three experiments to test the effectiveness of supplementing knowledge-bases with probabilistic models. The experiments are: (1) to predict the part-of-speech of highly ambiguous words, (2) to predict the intended interpretation of an utterance when more than one interpretation satisfies all known syntactic and semantic constraints, and (3) to learn case frame information for verbs from examples.

In the first experiment, the well-known *bi-Gram* and *tri-Gram* models [Church89] were used in part-of-speech tagging of highly ambiguous words in a sentence by probability. The probability required by the models was derived from a manually annotated corpus. Moreover, syntactic categories derived from a Machine Readable Dictionary (MRD) was used to cope with words that were missed from the training corpus.

In the second experiment, a *context-free* model was used to estimate the probability of each grammar rule. The grammar rules combined syntax and semantics in a unification formalism. Given the input string, the probability of a syntactic structure S was modeled by the product of the probabilities of the rules used in S. Testing with a tailored corpus in a personnel question-and-answer domain indicated that the model had the ability to learn the lexical, syntactic, and semantic features of unknown words from the context.

In the final experiment, Weischedel et al. attempted to acquire semantic knowledge for determining which phrases containing a particular verb or noun was meaningful. The knowledge was represented in *case frames* [Berwick89]. The knowledge acquisition procedure included: (1) bootstrapping by manually annotating each noun, verb, and proper noun in a given sample with the semantic class corresponding to it in the domain model, (2) based on the syntactic category provided by the annotated corpus and relationships between the lexical items and the concepts stated in step 1, statistically generalize the minimum data for discriminating cases for

attaching phrases to their head of the training set (i.e. the starting noun or verb), and (3) estimate the probability.

Overall, the experiments performed by Weischedel et al. were extensive. They covered a wide range of knowledge required in typical NLP applications, including: part-of-speech tagging, grammar rule generation and semantic analysis. Analysis techniques employed in the experiment were based on the simple probabilistic model common used in corpus-based researches. The major limitation of the approaches proposed in the experiment was that they all assume the corpus contained appropriate markers to indicate the existence of the required knowledge. However, the authors did not suggest any method to improve the efficiency of the corpus tagging process.

Similar to the last experiment, Pustejovsky et al. [Pustejovsky93] proposed an approach for the acquisition of lexical information for several classes of "nominal". The lexicon information was defined in the general framework of a generic lexicon structure as outlined in Pustejovsky (1991). Instead of tagging the corpus manually, Pustejovsky et al. suggested an external bootstrapping procedure to simplify the process. The bootstrapping procedure involved the parsing of word definitions in the Oxford Advanced Learners Dictionary and LDOCE to generate the initial lexical structures of 25000 words. This basic knowledge captured in the initial lexical structures was then enhanced with knowledge gleaned from a corpus. The corpus was pre-processed to build a database of phrasal relationships and then followed by performing sub-language analyses to acquire the required knowledge. The analyses

included: noun compound recognition and bracketing, generation of taxonomic relationship on the basis of collocation information, and extraction of information relating to noun's *qualia structure* [Pustejovsky93]. Similar to Pustejovsky, Zernik U. [Zernik91] proposed to use a dictionary to bootstrap his corpus based word-sense tagging algorithm for improving information retrieval accuracy.

Jacobs [Jacobs94] also proposed an alternative approach to eliminate the need of tagging the corpus in order to get word sense information from corpus for a multilingual text interpretation system SHOGUN. SHOGUN contained two types of knowledge: a core ontology of about 1000 concepts that supported word senses in the scored English and Japanese lexicons and a domain-specific knowledge base. The domain-specific knowledge base consisted of groups of words and phrases, for both of English and Japanese, extracted from the corpus. A two-stage process was designed for developing these word groupings. Major groupings were defined manually at first. It was then further expanded based on the corpus analysis. The corpus analysis consisted of two steps: (1) the expansion of manually defined word classes using weighted mutual information statistics, and then (2) the identification of words closely related to the groups again based on the mutual information statistics. Without the need to pre-process the corpus manually, expansion of the system by incorporating more corpus is much more cost effective.

In fact, a raw corpus (i.e. a corpus without pre-processing) can provide certain limited knowledge such as mutual information between words. In a comparison study, Niwa and Nitta [Niwa94] used the knowledge from a large corpus to

disambiguate word senses. Simple co-occurrence statistics were collected from the corpus for resolving the sense of the ambiguous words in a sentence using the method proposed by Wilks et al. The co-occurrence likelihood between any two words was calculated using the mutual information estimates proposed by Church and Hanks [Church89]. The co-occurrence likelihood was a probability indicating the chances of two words appears as neighbors. In the study, the neighborhood of a word was defined as 50 words before and after it. Unfortunately, the NLP application that can be handled with pure co-occurrence likelihood is rare.

Application of corpus in Chinese NLP is relatively new. In recent years, major study in Chinese corpus-based NLP focused on the acquisition of syntactical knowledge for Chinese parsers. A team of NLP researchers has designed a corpus-based Chinese parser that uses frame-based grammar - *Dependency Grammar* [Pan91, Huang92a, Huang92b, Yuan93]. A small size corpus was collected and manually annotated with syntactic makers. Collocation frequency on word pairs are computed directly from the corpus. It is used to train the parser. Since frame-based parser is very fine grain, its major drawback is that the training corpus has to be very large for its coverage to be acceptable. However, due to the need to manually tagging the corpus with syntactic maker, it is hard to build a large corpus.

To alleviate this problem, Yuan et al. [Yuan94] proposed to integrate the corpus based and frame-based techniques to form the RUSTIE parser. Unlike a frame-based parser which is suitable for infrequent grammatical structures, a rule-

based parser is coarse grain and is more suitable for high frequency. RUSTIE was designed to take advantage of both.

To develop a Chinese electronic dictionary, Chen K.J. [Chen94] endeavored to acquire the lexical knowledge from tagged corpora. The lexical knowledge of the electronic dictionary contained information about the syntactic rules and semantic preference relationship between words - this is often referred to as a lexicon. The semantic preference relationship comprises numeric data for representing the association strength between words in the tagged corpora. In order to reduce the man-power required to annotate a corpus, Chen proposed an iterative lexical knowledge acquisition procedure. Initially, lexicographers used a machine-assisted tool to build the basic lexical information for pre-processing the corpus. The pre-processed corpus was then further tagged manually with semantic makers to produce a tree-bank (i.e. corpus with semantic makers). Linguistic knowledge and selective association statistics were extracted from the corpus using the *n-Gram* method. They were used to enhance the basic lexical knowledge provided by the lexicographers. By repeating the above process, a complete lexicon was built. The iterative knowledge acquisition approach can only reduce the complexity of the problem but not completely solve it.

In general, a corpus contains many types of knowledge applicable to NLP systems and its analysis is straight-forward. But it cannot be used directly. It must be pre-processed in order to make the knowledge explicit for subsequent analysis. The type of pre-processing depends on the required knowledge. However, corpus

pre-processing usually involves a lot of man power and this limits the widespread acceptance of corpus-based NLP. Another major disadvantage of corpus-based approach is that the knowledge source (i.e. the corpus) in general has to be very large in order to achieve wide coverage. This explains why many researchers prefer linguistic-based techniques for knowledge acquisition.

2.3 Linguistic-Based Knowledge Acquisition

Different from the printed dictionary, a Machine Readable Dictionary (MRD) such as LDOCE explicitly provides some additional semantic information other than just meaning of words. This includes *box code* and *domain code* which are useful to NLP tasks. Box code contains approximately 25 semantic features of the word e.g. ABSTRACT, HUMAN, SOLID, etc. Domain code contains 2-level hierarchy of around 300 pragmatic code e.g., MEDICINE, BIOLOGY, LAW, etc. In the past, this information was only provided by standard printed English thesauri e.g. Roget's thesaurus [Roget88]. Now, with the semantic information on-line, researchers in English NLP can use LDOCE to directly solve many linguistic problems e.g. word sense disambiguation [Harder93]. Unfortunately, the linguistic resources available in Chinese is limited. The development of a comprehensive MRD is only slowly on its way [Chen94].

2.3.1 Knowledge Acquisition from Standard Dictionaries

In the past few years, many researchers have exploited the use of various information provided by a MRD in NLP. Amongst them, Wilks et al. [Wilks93] conducted the most extensive studies on transforming a MRD into a machine

tractable format (MTD) (i.e. a format usable for NLP). In a recent paper, Wilks et al. have proposed three different but related computational methods to transform the LDOCE into MTD. These methods differed in the amount of knowledge they started with and the kinds of knowledge they provided.

Method 1 - Statistics of co-occurrence of words in the word definition were collected from LDOCE. The co-occurrence statistics were then used to derive the relatedness of words using a function (i.e. *relatedness function*). Different relatedness functions produced different mappings between the co-occurrence statistics and the relatedness of words. It was used to represent different assumption on the relationship between co-occurrence statistics and semantic relationships of words. A total of six different relatedness functions were tested with the word sense disambiguation problem in the study. Since the method computed only the co-occurrence statistics, the bootstrapping data required in this method was little.

Method 2 - Based on the fact that definitions of words in LDOCE were written in about 2000 controlled vocabulary, Wilks et al. proposed to derive a natural set of semantic primitives from LDOCE and to use these primitives in the construction of an MTD. The construction procedure was divided into four steps: (1) determination of the defining senses in LDOCE[2], (2) derivation of a natural set of semantic primitives from LDOCE[3], (3) construction of a lexicon and a knowledge base using the natural set of semantic primitives, and (4) construction of a MTD by

[2] *Defining senses* are word senses used in the definitions of the controlled vocabulary.

[3] Semantic primitives are the subset of *defining senses* that are sufficient to represent all meanings of the controlled vocabulary.

means of bootstrapping from initial hand-crafted lexicon and knowledge bases prepared in step 3. The knowledge acquired, as a formalized set of definitions of sense entries, was defined in a nested predicate form, where the predicates are the semantic primitives. The main advantage of this method is that the MTD contains highly structured semantic information. But this is achieved at the cost of intensive manpower for constructing the basic lexicon and knowledge base.

Method 3 - In this method, a *chart parser* was developed to parse the word definitions and to produce phrase-structure trees with hand-crafted grammar. A phrase-structure tree produced by the parser was then interpreted by a tree interpreter to produce a frame-like structure. The tree interpreter employed hand coded semantic patterns and rules to perform pattern matching and inferencing in the analysis of the phrase-structure tree.

Without relying on the use of much external knowledge to bootstrap, Vossen et al. [Vossen89] followed the *stepwise lexical decomposition* framework of Dik [Dik78] and proposed to systematically store Meaning Descriptions (MDs) in LDOCE to form a semantic database. The basic methodology was to first apply an appropriate grammatical coding to the words of the controlled vocabulary and their inflected forms. This coding was then automatically inserted to all the MDs. A syntactic typology based on the major part-of-speech tags, e.g. nouns, verbs, and adjectives was then developed for each MD. This resulted a set of parser-grammar for the MDs. Applying these grammars to analyse the MDs led to syntactically identified premodifiers, kernels, postmodifiers, etc.

Different from the previous methods, Klavans et al. [Klavans93] focused on the understanding of the semantic structure of a frequently occurred head noun (*unit*) obtained from the dictionary definitions. He claimed that his method could be used in the development of an independent knowledge base for coping with the problem of sense mapping across dictionaries. In the experiment, word definitions from the Webster's Seventh New Collegiate Dictionary were parsed using the English Grammar namely PEG [Jensen93]. The parse trees were then converted into PROLOG terms.

Niwa and Nitta [Niwa94] tried to resolve word sense ambiguity using the knowledge from a standard dictionary. From the word definitions in the dictionary, a reference network of words was constructed to denote the distance between them. The network was a graph that showed which words were used in a word definition by giving a link between the word being defined and the words in the definition. The basic hypothesis of this approach was that if a word A was used in the definition of another word B, these words were expected to be strongly related. An algorithm was defined to estimate the distance between any two words in the reference network by taking into account of word frequency. A word vector was defined as the list of distances from a word to a certain set of selected words. It was used to disambiguate the sense of a word by using a method similar to the one proposed by Wilks et al [Wilks93]. The linguistic-based approaches discussed so far all work toward a goal of transforming MRDs into MTDs irrespective of the difference in the knowledge being captured.

Unlike the others, Binot and Jensen [Binot93] suggested that natural language was a kind of knowledge representation in nature and transformation of MRDs into MTDs could be avoided. Most knowledge can be directly expressed in natural language providing that the computers can access that information. In their experiment, they developed a semantic expert to use the knowledge on Webster's Seventh New Collegiate Dictionary to solve the prepositional phrase attachment ambiguities in a sentence. Definitions of words in the dictionary were parsed by PEG [Jensen93]. Pattern-matching was then performed on the parse tree to search all predefined patterns. The matched pattern was further expanded by using synonyms or taxonomy relations extracted from the dictionary. The expanded pattern was used to create a semantic network similar to that put forward by Calzolari and Zampolli [Calzolari90]. Furthermore, specific inference rules were provided for deducing relationships between words in cases where the dictionary entry failed to provide the necessary information.

In summary, a dictionary can also provide knowledge for different NLP applications. The major difference between the corpus-based and the linguistic-based knowledge acquisition approaches is that the latter required much less human involvement in the bootstrapping procedure. Furthermore, since the dictionary is domain independent, the knowledge extracted from it can provide a wider coverage than the same from a corpus.

2.3.2 Knowledge Acquisition from Standard Thesauri

Although MRD can provide abundant semantic information, Tsutsumi [Tsutsumi93] propounded a method for extracting semantic information from the taxonomy and synonym hierarchies developed by IBM T.J. Watson Research Center [Chodorow85]. This method was used to disambiguate multi-sense words in a sentence by using a number of testing sentences. A simple scoring scheme ranging from 0 to 100 was experimentally determined to indicate the semantic closeness between two words. For example, if two words were the same, then a score of 100 would be given; if a word was a synonym of the other, then it would be assigned with a smaller score e.g. 70; and if two words were under the same taxonomy, the score would even be lower e.g. 65.

In a related study, Tong et al. [Tong93] developed a system to automatically tag the sense of unknown compound words in Chinese running text. It was based on the knowledge provided by a MRD and the CILIN. One-syllable words and compound words and phrases formed by them were stored in the MRD. The compound words and phrases in the MRD contained implicit syntactic information useful to example-based reasoning about the senses of the Chinese words in context. Whenever an unknown word with an ambiguous character was encountered, the concept distance between the non-ambiguous part of it and the non-ambiguous part of the related compound words and phrases found in the MRD were derived from the CILIN. Similar to Tsutsumi, Tong et al. used three parameters to approximate the conceptual relatedness of any two words from the hierarchical structure in the CILIN.

Both Tsutsumi and Tong et al. used simple approximation to directly estimate the semantic relationships between semantic classes in the thesaurus based on the hierarchical relationships provided by it. Not until recently, Lua K.T. [Lua93a, Lua93c] has performed an extensive study on Chinese word semantic using the CILIN. He concluded that the CILIN provided rich knowledge for performing Chinese word semantic disambiguation. Furthermore, he also analysed in details the *semantic field* of the CILIN and claimed that the simple co-occurrence method could be used to measure the *associativeness* (i.e. conceptual distance or closeness) between different semantic categories [Lua93b].

2.4 Remarks

The knowledge contents in a thesaurus are clearer comparing to a dictionary. They comprise semantic relationship between different semantic classes. This is a useful form of knowledge for NLP. Unlike a dictionary, the structure of a thesaurus is more explicit. Knowledge acquisition from a thesaurus is simpler for it is not required to be pre-processed. In addition, it also has the property of domain independence. For these reasons, I propose to develop a linguistic-based knowledge acquisition method to acquire the inter-word semantic relationship from the widely studied Chinese thesaurus - CILIN《同義詞詞林》.

CHAPTER 3 A METHOD TO EXTRACT THE INTER-WORD SEMANTIC RELATIONSHIP FROM《同義詞詞林》

Our method to extract the inter-word semantic relationship from the CILIN 《同義詞詞林》involves several steps. First, an *association network* was designed in representing semantic knowledge. Second, a *semantic association model* was used to extract the required semantic information from the CILIN to build a knowledge base. Finally, an *inter-word semantic function* was formulated to derive the inter-word semantic relationship using the semantic knowledge captured in the knowledge base. In the following sections, I will first provide some background information and then describe each of the three steps in detail.

3.1 Background

In this section, I will first describe the structure of the CILIN, the knowledge representation of the machine tractable thesaurus, and a method of simple co-occurrence[4] proposed by Lua K.T [Lua93b] for extracting the semantic knowledge from the CILIN.

[4] The principle of the method proposed by Lua K.T. is the same as the simple co-occurrence method but the author did not use the term in his paper.

3.1.1 Structure of 《同義詞詞林》

The CILIN contains approximately 63,600 entries for about 52,500 Chinese words with each word having 1.21 entries on the average[5]. It classifies the words according to a three-level semantic tree structure. This hierarchical structure reflects the semantic relationship between words. It is defined by 12 major (top level), 95 medium (middle level), and 1428 minor (bottom level) semantic classes. Each minor semantic class in terms comprises a set of words. Effectively, words under the same minor semantic class share the concept of this class. Figure 1 is an example from the thesaurus. Refer to the figure, the hierarchical structure spans from left to right, i.e. A-L are the 12 major classes, Aa-An are examples of middle classes and Aa01-Aa06 are examples of minor classes. For example, the word 誰 (who) i.e. minor class Aa06, belongs to the middle class Aa (泛稱, common name) and the major class A (人, human). Further, refer to p.2 of the CILIN, the minor class Aa06 [誰] consists of 9 words, see below:

Aa06 [誰]

誰	孰	誰人	誰個	何人	何許人	哪人
哪位	若人					

A word may appear in more than one branch of the hierarchy. On the average, a word has 1.21 entries. Further a major class covers multiple middle classes and each middle class may in turn have multiple minor classes. Two

[5] The basic statistics of the CILIN are slightly different from that reported in the Lua K.T. 1993b paper. The major reason is that, in the computerization of the CILIN, we have converted the CILIN from the simplified character form to the traditional character form to meet the local (i.e. Hong Kong) requirements.

different words may have some overlapping in their semantic classes at different levels. The amount of overlapping in the semantic classes of two words reflects the similarity of the words. Lau K.T. has studied this in some detail in [Lau93a, Lau93c].

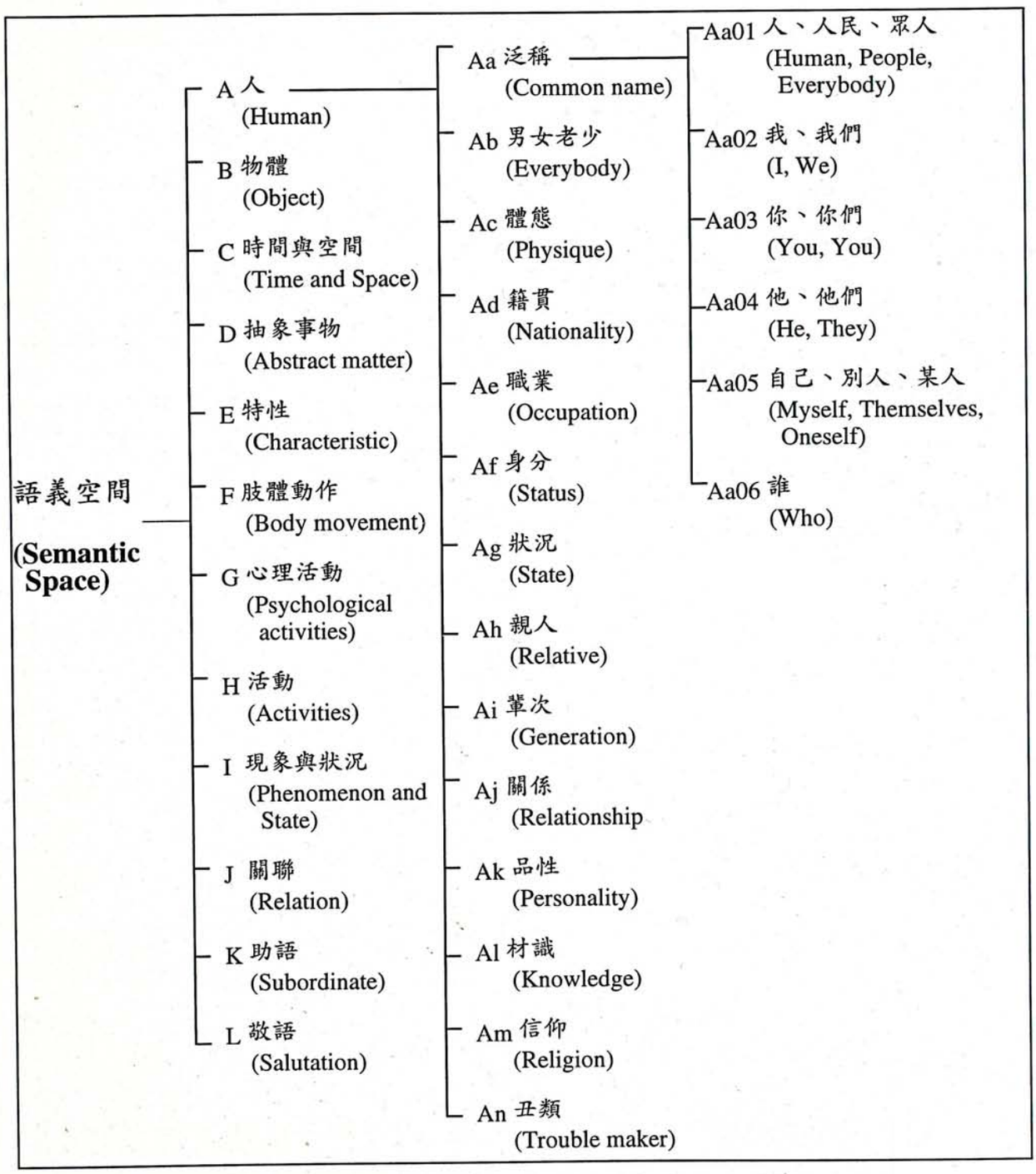


Figure 1 An Example from 《同義詞詞林》

3.1.2 Knowledge Representation of a Machine Tractable Thesaurus

A well constructed thesaurus has long been recognized as a valuable information resource for NLP applications particular in information retrieval [Crouch88, Kimoto90, Paice91, Crouch92, Lu94]. Lua and Yi [Lu94] recently proposed a thesaurus model to capture the semantic relationships among words and concepts. The model was targeted for multiple Chinese NLP applications. In their model, semantic knowledge of words was defined on the basis of concepts and semantic closeness between any two words. Relationships among concepts were defined in a frame structure with six relationship types represented as its slots. In addition, a *weighting scheme* was designed to represent the closeness between the concepts in each relationship. A weight in the range of 0 to 1 was used to quantify the relevance between directly linked concepts. Furthermore, an algorithm was defined for computing the relevance of the indirectly linked concepts through the associated directly linked ones.

Although Lua and Yi provided a good framework in the construction of a machine tractable thesaurus, they did not address the knowledge acquisition problem. Referring to Lua and Yi's thesaurus, it is hard, even for a linguist, to directly assign an appropriate number to indicate the relevance of the semantic relationship (i.e. the weight).

3.1.3 Extracting the Semantic Knowledge by Simple Co-occurrence

Traditionally, a thesaurus represents the knowledge about word semantics. It groups the closely related words together to form *semantic classes* and each class is further classified into a tree structure according to their level of abstraction. A

conventional thesaurus is designed for human readers. Human can follow the links of a tree to search for a word based on a given semantic or to start from a word of known semantic to search for alternative words with related meaning. Although each link of the tree provides information about the hierarchical relationship, a thesaurus does not quantify the weight of this link. Without any inter-link information, even if computers were employed to navigate this type of tree, a user would not know how to begin. This is because when a thesaurus is made, lexicographers assumed that the users would possess some basic knowledge about the relationships between the semantic classes. For instance, among the semantic classes Hc10:控制(control), Je09:支配 (in-charge), and Kb06:依靠 (depend) in the CILIN[6], the user would know that the classes 控制 was closer to 支配 than 依靠 even though they were classified under different sub-trees in the CILIN. In reality, this type of knowledge is well embedded in the thesaurus and can be extracted fully and automatically.

In a similar work, Lua K.T. [Lua93b] studied the semantic field of the CILIN and attempted to re-classify the major classes of the CILIN using the theory in psychological linguistic. In the study, he pointed out that the classification scheme of the CILIN was based the subjective decision of theoretical linguists. Thus, the number of hierarchical levels of the CILIN was devised for the purpose of convenience. Since many Chinese words have multiple entries in different semantic classes, he proposed to use co-occurrence statistics to derive the closeness (or proximity) for the major, middle, or minor semantic classes. Based on the

[6] In the CILIN, concept of a semantic class may be denoted by multiple words e.g. the class Hc10 is "控制，把持". In the example, only the first word "控制" is used.

observation in theoretical linguistics, he defined two coefficients, namely, the *lender* of the meaning (l_o) and the *recipient* of the meaning (l_i). These coefficients measured the degree in which a semantic class could accept another class and the degree in which a semantic class could accept by another class, respectively (see Figure 2). He computed such coefficients at the second level (i.e. middle classes) and concluded that the derived coefficients could reflect the semantic relationships between all semantic classes.

l_o and l_i are coefficients to measure the semantic relationship between the semantic classes from two different angles and, therefore, are asymmetric. To avoid using asymmetry, he further defined an *associativeness* coefficient, *A*, by taking the geometric means of l_o and l_i (i.e. $A = \sqrt{li * lo}$). He suggested to use *A* to represent the average semantic relationship between two semantic classes. In his paper, he applied associativeness coefficients to few psycho-linguistic experiments and concluded that such coefficients were significant and practical.

At a first glance, from the conclusion of Lua K.T., it seems that the conventional simple co-occurrence method is effective for measuring the semantic relationship between different semantic classes and the associativeness coefficient seems to be a good index to denote their semantic relatedness. However, with a multiplicity of 1.2 entries in a word, the overlapping between any two semantic classes is very small. From this observation, the significance of co-occurrence statistics suggested by Lua K.T is skeptical. Bearing this skepticism in mind, I have repeated Lua's co-occurrence analysis on the CILIN. The results confirmed my

suspicion over the significance of the co-occurrence measures for the CILIN. In Section 3.3.1, I will describe this problem in more details and discuss the findings from the repeated Lua' s experiment.

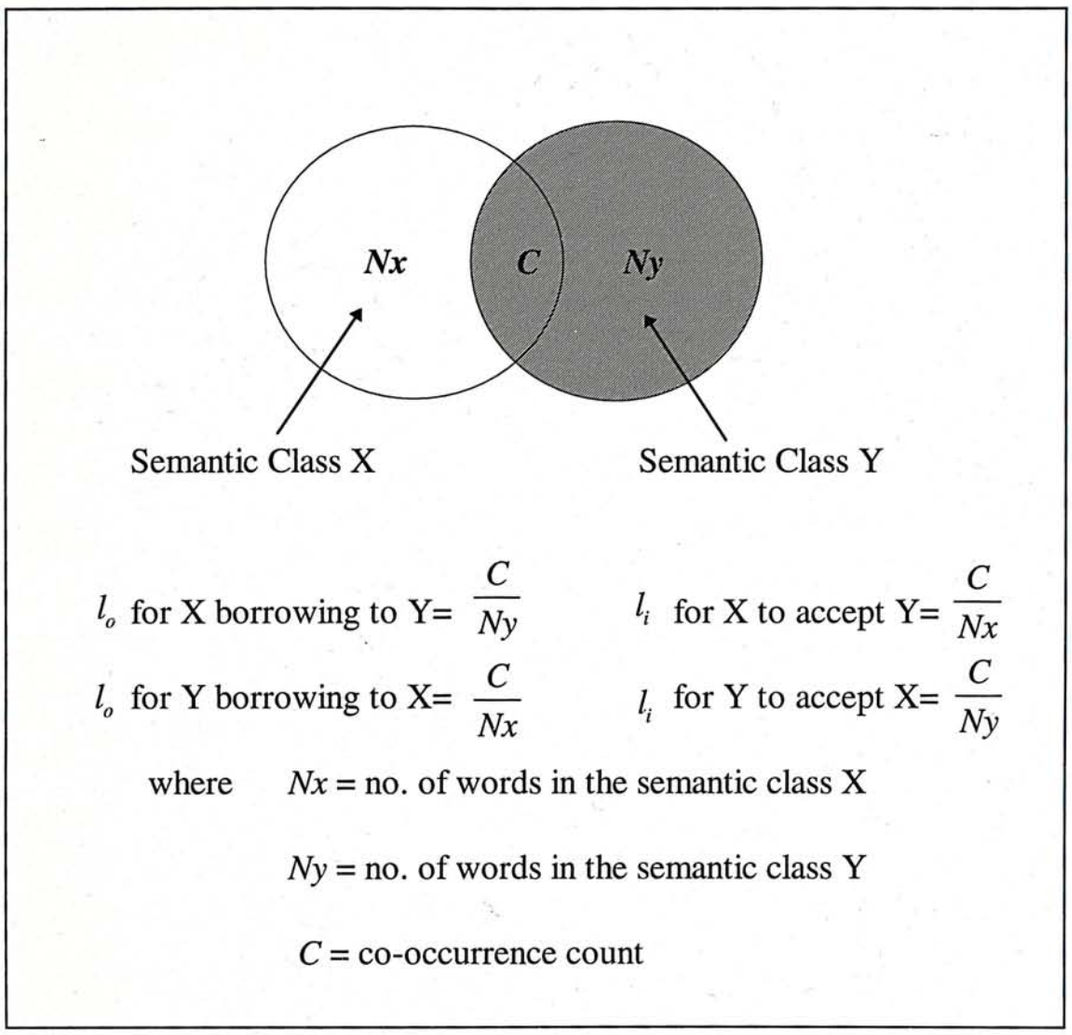


Figure 2 Definitions of Lender and Recipient (l_o, l_i) [Lua93b]

3.2 Association Network

As stated in Section 3.1.3, the inter-link relationship between the semantic classes is well embedded in the CILIN and can be made explicit by knowledge

acquisition techniques. For this purpose, a new knowledge acquisition model - namely, the *semantic association model* is introduced.

I propose to use a massive *association network* (i.e. the knowledge structure for the semantic association model), similar to that discussed in Section 3.1.2, to capture the semantic relationships between different semantic classes. Nodes of the association network represent the semantic classes and the connection between them are quantified with a numeric value ranging from 0 to 1 (i.e. *connection weight*) to represent their semantic closeness as shown in Figure 3 (refer to the figure, Wij is the connection weight). In this model, the higher the connection weight is, the closer the semantic relationship between two nodes becomes. A value of 1 means perfect relationship and a value of 0 means total unrelatedness. For example, the connection weight between the semantic classes 控制 and 支配 is assigned with a higher connection weight (0.27) than the same between 控制 and 依靠 which is 0.09 (see Figure 4). Under the association network, semantic relationships between semantic classes are explicitly defined by the connection weights and are easily accessible. Moreover, unlike Lua K.T.'s method, which could only work out the semantic relationships for semantic classes on one level, my approach can derive the connection weights of semantic classes in all 3 semantic levels of the CILIN. Therefore, it is more comprehensive.

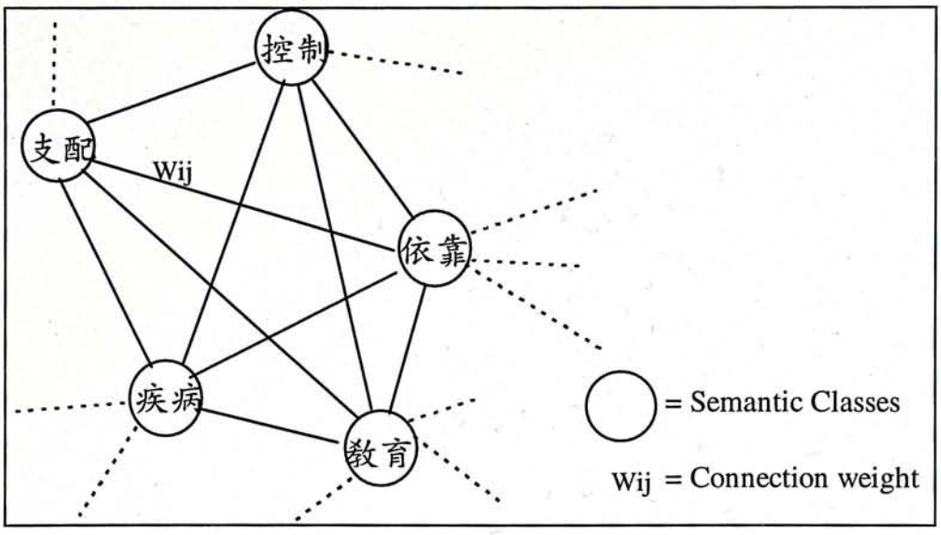


Figure 3 An Example of Association Network

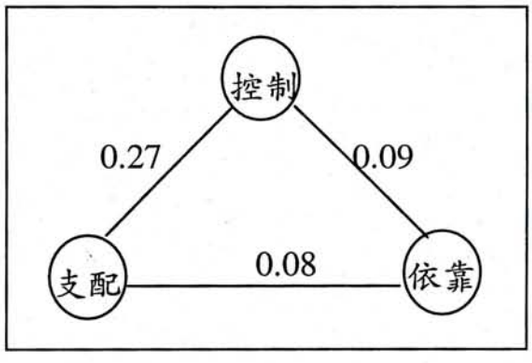


Figure 4 An Example of Connection Weight

3.3 Semantic Association Model

A fully automatic knowledge acquisition model is described in this section. It is used to gather the required knowledge from the thesaurus (i.e. the CILIN) for the construction of the association network described before. The semantic association model is bootstrapped internally. The objective of the model is to transform the

CILIN from a hierarchical tree structure to the association network as illustrated in Figure 5.

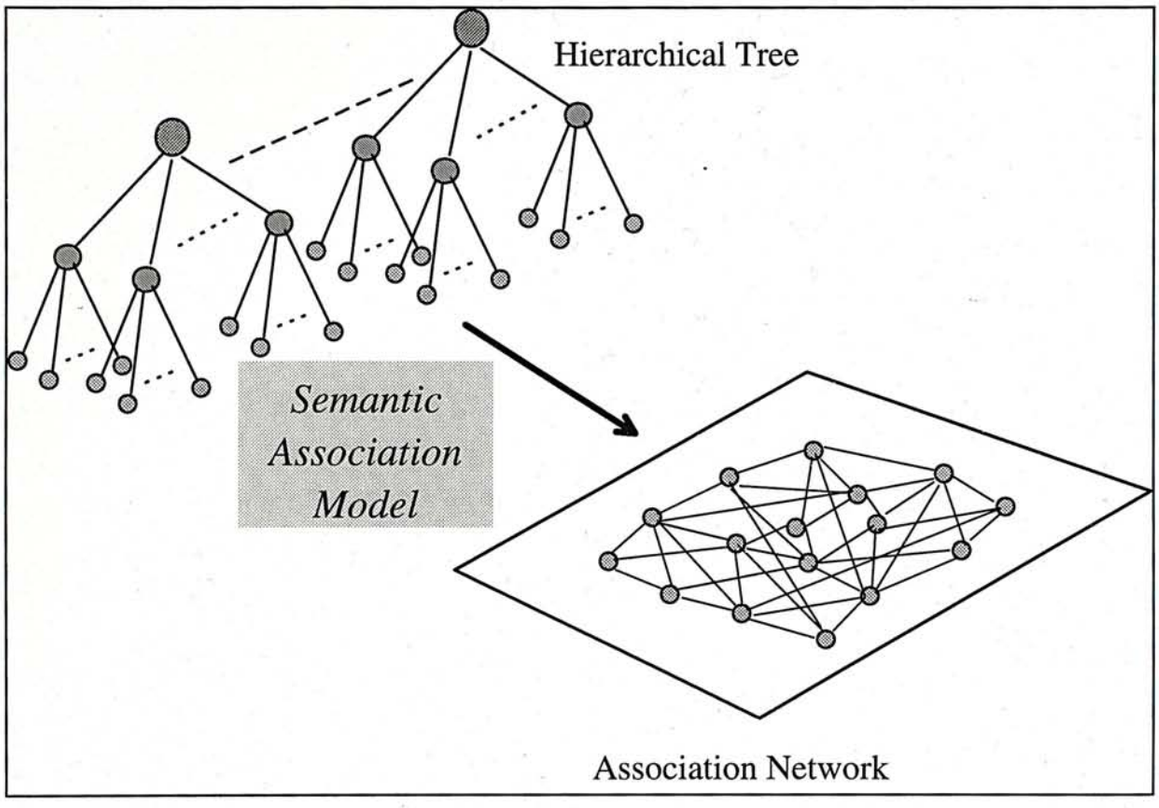


Figure 5 Purpose of Semantic Association Model

3.3.1 Problems with the Simple Co-occurrence Method

Intuitively, the simple co-occurrence method, a versatile approach in corpus analysis, seemed to be an effective algorithm for extracting semantic knowledge from the CILIN as suggested by Lua K.T. (refer to Section 3.1.3). The associativeness coefficient (*A*) defined by Lua could be used as the connection weight of the association network. However, the basic statistics of the CILIN show that on the average a word in the CILIN has only about 1.2 entries. It means that each semantic class on the average can have 20% of its words shared by the others. However, the

number of possible pairs that could be formed at the middle semantic level had a total of 8742 (94*94 - 94) pairs. If the words were evenly distributed, then the overlapping between any 2 middle semantic classes would only be approximately 0.2% (20% divided by 93) and, therefore, the co-occurrence value would be as small as 0.002. In the other extreme, if the distribution was biased and all words in one semantic class was completely duplicated in another one, the simple co-occurrence value would be 0.2 in 94 pairs but 0 for the remaining 8648 pairs. Based on this simple estimation, the simple co-occurrence approach can hardly produce the association coefficient values as high as the ones reported in [Lua93b]. To clarify this, I have constructed a machine readable CILIN and repeated the Lua K.T.'s experiment on the whole thesaurus. Table 1 is a comparison of the results from my experiment with those given in "Table 1" in Lua's paper. In the table, the magnitude of the co-occurrence counts computed in the repeated experiment is tally with that produced by Lua. The figures about the number of words in the semantic classes vary greatly, especially for Np (i.e. columns 2 & 7 and 3 & 8). The word count reported by Lua is much smaller than the one I computed. It seems unreasonable to have so many classes containing less than 10 words because the second level semantic classes have over 600 words on the average. Due to the significant difference in word count, the lender and recipient coefficients calculated in my experiment is much smaller than Lua's. My experiment found that the lender coefficients in most cases were less than 0.1 rather than all over 0.3 in Lua's case.

In addition, I also computed the associative coefficient of the middle level semantic classes. The result is shown in Table 2. It was found that only 22 out of

the 8742 links between the 94 semantic classes are greater than 0.05. With the majority of the coefficients below 0.05, there is not much one can do with the coefficients in NLP. Thus, it is concluded that the simple co-occurrence method cannot effectively measure the semantic relationships between semantic classes in the CILIN.

There are three major problems with the simple co-occurrence method for extraction of semantic information from the CILIN.

(1) The simple co-occurrence method can only identify the *strong semantic relationship* from the semantic classes (e.g. those links with $A > 0.05$).

(2) The semantic classes was well-clustered by the lexicographer into a number of basic concepts but these clusters have little connections among themselves.

(3) The conventional simple co-occurrence method uses exact-matching to compute the semantic relatedness between groups of words. However, exact-matching works only at the word level. This is reasonable and realistic only if words are semantically unique. But, this is not always true in practice.

Table 1 Comparing the Result of the Repeated Experiment against that Reported by Lua K.T.

Semantic Class (Second level)		Results Reported by Lua K.T. [Lua93b - Table 1]					Results of the Repeated Experiment				
p	q	(1) Cpq	(2) Np	(3) Nq	(4) lo	(5) li	(6) Cpq	(7) Np	(8) Nq	(9) lo	(10) li
Ak	Ed	4	6	415	0.666	0.010	4	343	2993	0.012	0.001
Hh	Fa	20	36	488	0.555	0.040	16	280	745	0.057	0.021
Kc	Ka	30	57	285	0.526	0.105	38	198	1184	0.192	0.032
Ac	Bh	2	4	273	0.500	0.007	2	113	1256	0.018	0.002
Al	Ed	6	12	415	0.500	0.014	11	406	2993	0.027	0.004
Ab	Ah	15	33	80	0.454	0.187	61	289	864	0.211	0.071
Aj	Ed	9	21	415	0.428	0.021	11	496	2993	0.022	0.004
Hb	Fa	24	56	488	0.428	0.049	16	459	745	0.035	0.021
Bc	Bk	30	73	222	0.410	0.135	35	131	1008	0.267	0.035
Dh	Ed	9	22	415	0.409	0.021	8	241	2993	0.033	0.003
Kc	Kb	23	57	104	0.403	0.221	17	198	144	0.086	0.118
Am	Ed	2	5	415	0.400	0.004	2	100	2993	0.020	0.001
Ae	Ed	14	35	415	0.400	0.004	10	957	2993	0.010	0.003
Hk	Bn	2	5	164	0.400	0.012	2	77	1033	0.026	0.002
Ec	Ed	33	84	415	0.392	0.079	30	424	2993	0.071	0.010
Jc	Hj	11	29	364	0.379	0.030	8	144	2332	0.056	0.003
Id	Fa	89	237	488	0.375	0.182	81	712	745	0.114	0.109
Ai	Ah	3	8	80	0.375	0.037	17	110	864	0.155	0.020
Ai	Ca	3	8	128	0.375	0.023	3	110	1225	0.027	0.002
Jb	Ed	18	48	415	0.375	0.043	22	262	2993	0.084	0.007
Kb	Hj	38	104	364	0.365	0.104	27	144	2332	0.188	0.012
Bb	Dn	20	55	301	0.361	0.035	16	85	734	0.188	0.022
Hh	Hj	13	36	364	0.361	0.035	14	280	2332	0.050	0.006
Hf	Fa	18	50	488	0.360	0.036	11	254	745	0.043	0.015
Hd	Fa	42	121	488	0.347	0.086	28	597	745	0.047	0.038
Gc	Ed	9	26	415	0.346	0.021	11	100	2993	0.110	0.004
Ec	Eb	29	84	358	0.345	0.081	40	424	1867	0.094	0.021
Ea	Eb	39	113	358	0.345	0.108	52	462	1867	0.113	0.028

[N.B. (a) Cpq = Co-occurrence count between Semantic Classes p & q, (b) Np = No. of words in Class p, (c) Nq = No. of words in Class q, (d) lo = Lender, and (e) li = Recipent]

Table 2 Semantic Classes (Second Level) with *Associativeness* > 0.05

Semantic Classes (Second Level)		Lender	Recipient	Associativeness
p	q	li	lo	$A = \sqrt{li * lo}$
Ab	Ah	0.211	0.071	0.122
Fa	Id	0.109	0.114	0.111
Kb	Kc	0.118	0.086	0.101
Bc	Bk	0.267	0.035	0.096
Eb	Ed	0.117	0.073	0.093
Ka	Kc	0.032	0.192	0.078
Bb	Bc	0.094	0.061	0.076
Bb	Dn	0.188	0.022	0.064
Ed	Ka	0.039	0.098	0.062
Kd	Ke	0.058	0.065	0.061
Ed	Ee	0.055	0.065	0.060
Gc	Jc	0.070	0.049	0.058
Hf	Kb	0.043	0.076	0.058
Bb	Fa	0.165	0.019	0.056
Ea	Eb	0.113	0.028	0.056
Ah	Ai	0.020	0.155	0.055
Aj	Db	0.048	0.061	0.054
Dn	Fa	0.054	0.054	0.054
Ja	Kc	0.067	0.040	0.052
Je	Kb	0.030	0.090	0.052
Bc	Kb	0.053	0.049	0.051
Fa	Hi	0.099	0.026	0.051

Semantic overlapping between words cannot simply be determined by a exact-matching. A word different from the others can be partially related in semantics. For instance, the words 荧光屏 (screen) and 視力 (vision) are semantically related. This indirect semantic relationship or weak relationship (i.e. *weak semantic relationship*) cannot be obtained at the word level. Therefore, extraction of the semantic relationships from the CILIN using simple co-occurrence

statistics is partial and incomplete. Based on this observation, our new approach derives the semantic relationships on semantic level rather than on word level.

3.3.2 Methodology of Semantic Association Model

Our new model derives a semantic relationship between two semantic classes by computing the semantic association among them rather than counting the number of co-existing words (i.e. simple co-occurrence statistics between them). It can be easily proved that the new model is in fact a superset of the simple co-occurrence .

A word is the basic unit to represent semantics. It is effectively a container which often has multiple semantics. Theoretically, the lexicographer summarizes this knowledge in the thesaurus by clustering words into semantic classes. The semantic classes assigned to a word reflect the meaning of this word to the lexicographer. On the other hand, lexicography in turn uses words to explain the semantic classes defined by him/her. Therefore, words and semantic classes have a cyclic mutual relationship.

In the new model, I propose to derive the association between two semantic classes SC_x and SC_y based on semantic information. The concept of this association is depicted in Figure 6. Referring to the figure, the semantic of the words in class SC_y are projected to form the Semantic Space SP on the semantic plan. The association between the word W_x and the class SC_y is then determined by performing a similar projection of W_x on the plan. The semantic class which W_x belongs to is excluded. The overlapping between the two projections is then

determined. This process of projection and overlap determination is repeated for every classes to which W_x belongs. After all the classes are processed, the total amount of overlapping is summed up to give the total semantic association score *SA*. Similar to the lender and recipient coefficients produced by Lua, *SA* is an asymmetric value (i.e. computation of *SA* from SC_x to SC_y is different from SC_y to SC_x), the geometric mean of these two values is taken to derive the average semantic relationship between the semantic classes.

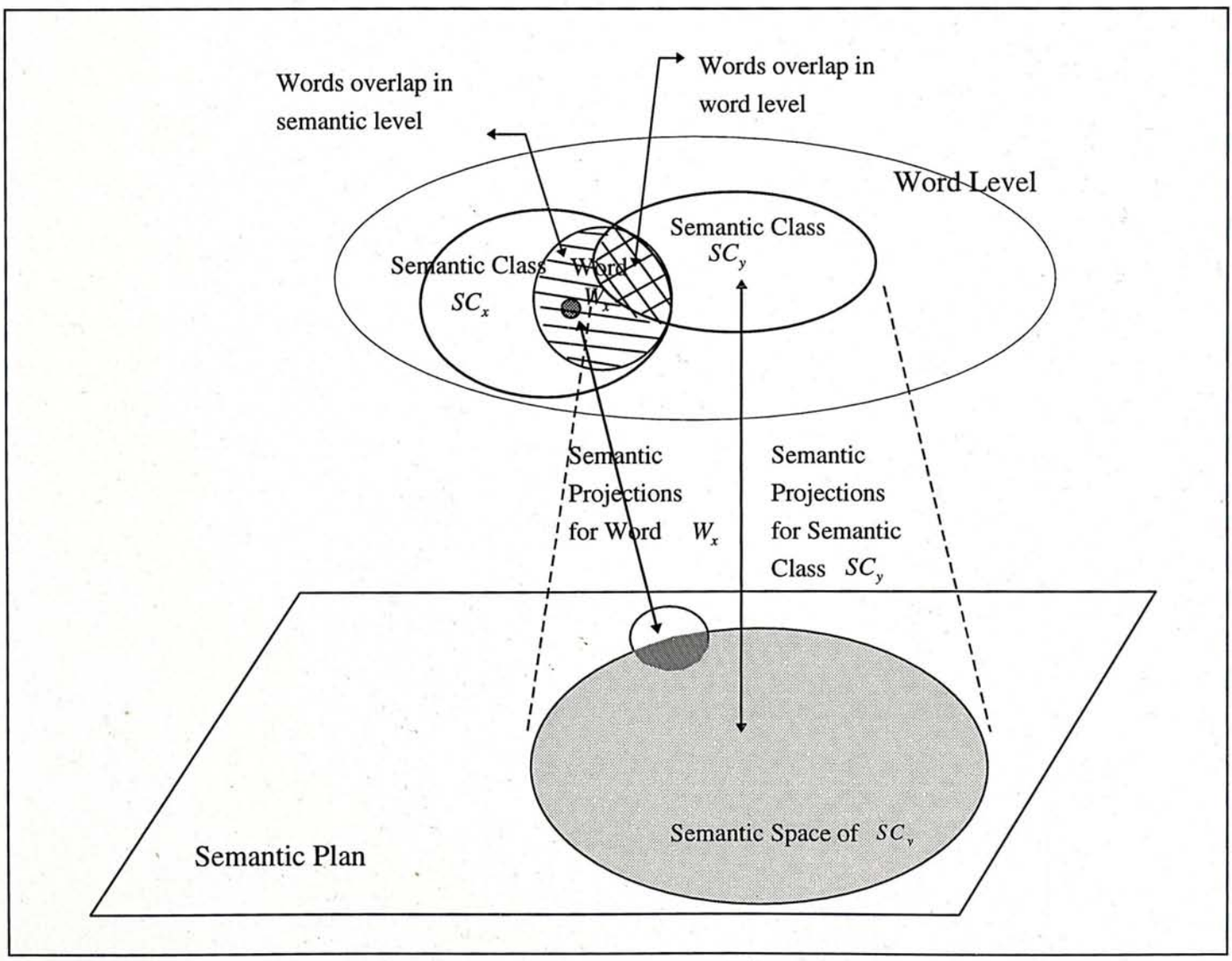


Figure 6 Methodology of the Association Network Model

3.3.2.1 Mathematical Formulation

The mathematical formulation follows exactly the model description. First, an association function, $Association(W_x, SC_y)$, is defined for determining the semantic association between a word W_x and a semantic class SC_y on the semantic plan as follows:

$$Association(W_x, SC_y) = \frac{1}{N_{wx}} \sum_{\forall sc_{wx} \neq SC_x} Cover(sc_{wx}, SC_y) \quad \textbf{(1)}$$

where

W_x is the word of semantic class SC_x.

SC_y is the semantic class.

N_{wx} is the number of semantic classes assigned to *W*.

sc_{wx} is a semantic class of the word W_x.

$Cover(sc_{wx}, SC_y)$ is a binary function to indicate if a semantic class of word *W* is covered by the semantic space SP_y of the semantic class SC_y and is defined as:

$$Cover(sc_w, SC_y) = \begin{cases} 1 & \text{if } sc_{wx} \in SP_y \\ 0 & \text{otherwise} \end{cases}$$

$$where \quad SP_y = \left\{ \sum sc_v \ \forall \text{ word } v \text{ in } SCy \right\}$$

The association function will return a value in the range of 0 to 1 with 1 to mean that all the concepts of the word W_x is covered by the words in the semantic class SC_y and 0 to mean that the semantic space of SC_y does not contain any concept of W_x. Clearly, if W_x exists in SC_y (i.e. W_x co-occur in SC_y), the semantic space of SC_y will contain all the semantic of W_x. For example, the semantic association between the word 源頭 (W_x), in the semantic class Db02 [根源; 來源] (SC_x), and the semantic class Ja05 [起源; 屬於] (SC_y) is derived as follows:

Example - Derivation of the semantic association between a word and a semantic class

Semantic class Db02 [根源; 來源] has 27 words:

本源	肥源	光源	能源	淵源	源頭活水
波源	根	河源	起源	源	資源
財源	根子	貨源	泉源	源頭	
電源	根苗	來源	熱源	源泉	
洞府	根源	濫觴	水源	源流	

Semantic class Ja05 [起源; 屬於] has 8 words:

導源	根源	起源	有賴
發源	來源	取決	源

Semantic classes of word 源頭 W_x in the semantic class Db02 [根源; 來源]:

Cb08, Db02

Semantic classes of the words in of Ja05 [起源; 屬於] includes:

Cb08, Db02, Ja05

Now, semantic space for Ja05 $SP_{Ja05} = \{\text{Cb08}, \text{Db02}, \text{Ja05}\}$

$N_w = 1$

$sc_w \neq \text{Db02} = \{\text{Cb08}\}$

Therefore,

Semantic association between W and Ja05

$$Association(W, \text{Ja05}) = \frac{1}{N_w} \sum_{\forall sc_w \neq \text{Db02}} Cover(sc_w, \text{Ja05})$$

$$= Cover(\text{Cb08}, \text{Ja05})$$

$$= 1 \qquad \square$$

Next, the relatedness function, $Relatedness(SC_x, SC_y)$, is defined to derive the semantic relationship between semantic classes SC_x and SC_y from the semantic association of the individual words in the classes as follows:

$$Relatedness(SC_x, SC_y) = \sqrt{\frac{SA(x, y)}{|SC_x|} \times \frac{SA(y, x)}{|SC_y|}} \qquad \textbf{(2)}$$

where

$|SC_x|$ and $|SC_y|$ are the number of words of semantic classes SC_x and SC_y respectively.

$SA(x, y)$ is the semantic association from the semantic class x to semantic class y; similarly $SA(y, x)$ is the same from class y to class x. They are defined as follows:

$$\boxed{\begin{aligned} SA(x,y) &= \sum_{\forall W_x \text{in} SC_x} Association(W_x, SC_y) \\ SA(y,x) &= \sum_{\forall W_y \text{in} SC_y} Association(W_y, SC_x) \end{aligned}} \quad (3)$$

For the two semantic classes used in the example in Section 3.3.2.1, the relatedness value is derived as below:

Example - Computation of the relatedness of two semantic classes

Follow the steps in Section 3.3.2.1,

Semantic association from the semantic Db02 to semantic Ja05

$$SA(\text{Db02}, \text{Ja05}) = \sum_{\forall W \text{in} \text{Db02}} Association(W, \text{Ja05})$$
$$= 5$$

Similarly, semantic association from the semantic Ja05 to semantic Db02

$$SA(\text{Ja05}, \text{Db02}) = \sum_{\forall W \text{in} \text{Ja05}} Association(W, \text{Db02})$$
$$= 4$$

And, $|\text{Db02}| = 27$
$|\text{Ja05}| = 8$

Therefore, relatedness between the semantic classes Db02 and Ja05

$$Relatedness(\text{Db02}, \text{Ja05}) = \sqrt{\frac{SA(\text{Db02}, \text{Ja05})}{|\text{Db02}|} \times \frac{SA(\text{Ja05}, \text{Db02})}{|\text{Ja05}|}}$$
$$= \sqrt{\frac{5}{27} \times \frac{4}{8}}$$
$$= 0.3043 \quad \square$$

The relatedness function (2) uses the geometric means to compute the average semantic relatedness between the semantic classes. Similar to the association function, it returns a value ranging from 0 to 1 to denote the semantic relatedness between two semantic classes with higher value being assigned to closer semantic relationship. By definition, the function is symmetric (i.e. $Relatedness(SC_x, SC_y) = Relatedness(SC_y, SC_x)$) and therefore it is a non-directional measure for the semantic relatedness between two semantic classes. By applying the relatedness function, we can prove that this model gives a value of 1 for two equal semantic classes (the simple co-occurrence statistics for the same is also 1). This implies that the value return by the relatedness function in Equation (2) is a normalized figure with a maximum value 1 in the case where two semantic classes coincide. Having this property, the relatedness function can be used to indicate the relative closeness (in semantic) between the semantic classes. The proof is shown below:

Proof

For two semantic classes be $SC_x = SC_y = SC$,

$$\begin{cases} Association(W, SC_x) = 1 & \forall W \in SC_x \\ Association(W, SC_y) = 1 & \forall W \in SC_y \end{cases}$$

$$\Rightarrow \begin{cases} SA(x,y) = \sum_{\forall W_x \text{in} SC_x} Association(W_x, SC_y) = |SC| \\ SA(y,x) = \sum_{\forall W_y \text{in} SC_y} Association(W_y, SC_x) = |SC| \end{cases}$$

Therefore,

$$Relatedness(SC_x, SC_y) = \sqrt{\frac{SA(x,y)}{|SC_x|} \times \frac{SA(y,x)}{|SC_y|}}$$
$$= \sqrt{\frac{|SC|}{|SC|} \times \frac{|SC|}{|SC|}}$$
$$= 1$$

Hence

$$Relatedness(SC, SC) = 1 \qquad \square$$

Thus, the relatedness function can be used to compute the connection weight of the association network.

3.3.2.2 Experimental Results

The semantic association model has been implemented and tested with the CILIN. The semantic relationships (i.e. connection weights) between the semantic classes were extracted and stored in a database using the mathematical models described in Section 3.3.2.1. The knowledge about the semantic relationships is now in a format ready for determining the inter-word semantic relationship. An experiment has been conducted to compare the efficiency of the semantic association model against the simple co-occurrence model in deriving the semantic association between the major (i.e. top level) semantic classes of the CILIN. Table 3 and Table 4 are the figures given by the simple co-occurrence model and the association network model respectively. For the reasons described in Section 3.3.1, the simple co-occurrence method can only produce a coefficient ranging from 0 to 0.079 but the semantic association network gives values ranging from 0.015 to 0.246 which are significantly higher.

Table 3 Using Simple Co-occurrence Model to Derive Semantic Association

	A	B	C	D	E	F	G	H	I	J	K	L	Total
A	-	0.016	0.012	0.038	0.020	0.001	0.002	0.016	0.004	0.011	0.011	0.000	0.13
B	0.016	-	0.021	0.049	0.018	0.028	0.005	0.020	0.014	0.012	0.013	0.000	0.20
C	0.012	0.021	-	0.040	0.022	0.008	0.003	0.010	0.014	0.008	0.021	0.002	0.16
D	0.038	0.049	0.040	-	0.040	0.023	0.026	0.051	0.030	0.035	0.034	0.004	0.37
E	0.020	0.018	0.022	0.040	-	0.020	0.061	0.040	0.062	0.038	0.059	0.009	0.39
F	0.001	0.028	0.008	0.023	0.020	-	0.026	0.079	0.061	0.044	0.036	0.002	0.33
G	0.002	0.005	0.003	0.026	0.061	0.026	-	0.033	0.030	0.026	0.023	0.004	0.24
H	0.016	0.020	0.010	0.051	0.040	0.079	0.033	-	0.063	0.060	0.042	0.021	0.44
I	0.004	0.014	0.014	0.030	0.062	0.061	0.030	0.063	-	0.051	0.042	0.007	0.38
J	0.011	0.012	0.008	0.035	0.038	0.044	0.026	0.060	0.051	-	0.054	0.002	0.34
K	0.011	0.013	0.021	0.034	0.059	0.036	0.023	0.042	0.042	0.054	-	0.004	0.34
L	0.000	0.000	0.002	0.004	0.009	0.002	0.004	0.021	0.007	0.002	0.004	-	0.06
Total	0.13	0.20	0.16	0.37	0.39	0.33	0.24	0.44	0.38	0.34	0.34	0.06	3.36

Table 4 Using Semantic Association Model to Derive Semantic Association

	A	B	C	D	E	F	G	H	I	J	K	L	Total
A	-	0.070	0.079	0.098	0.092	0.076	0.068	0.086	0.084	0.102	0.111	0.015	0.88
B	0.070	-	0.084	0.111	0.103	0.118	0.084	0.105	0.107	0.116	0.116	0.015	1.03
C	0.079	0.084	-	0.124	0.108	0.100	0.082	0.102	0.117	0.126	0.129	0.017	1.07
D	0.098	0.111	0.124	-	0.162	0.170	0.139	0.160	0.167	0.186	0.182	0.037	1.54
E	0.092	0.103	0.108	0.162	-	0.171	0.160	0.164	0.187	0.189	0.199	0.059	1.59
F	0.076	0.118	0.100	0.170	0.171	-	0.167	0.206	0.223	0.246	0.221	0.051	1.75
G	0.068	0.084	0.082	0.139	0.160	0.167	-	0.151	0.179	0.176	0.183	0.062	1.45
H	0.086	0.105	0.102	0.160	0.164	0.206	0.151	-	0.186	0.205	0.195	0.050	1.61
I	0.084	0.107	0.117	0.167	0.187	0.223	0.179	0.186	-	0.233	0.217	0.056	1.76
J	0.102	0.116	0.126	0.186	0.189	0.246	0.176	0.205	0.233	-	0.244	0.040	1.86
K	0.111	0.116	0.129	0.182	0.199	0.221	0.183	0.195	0.217	0.244	-	0.052	1.85
L	0.015	0.015	0.017	0.037	0.059	0.051	0.062	0.050	0.056	0.040	0.052	-	0.45
Total	0.88	1.03	1.07	1.54	1.59	1.75	1.45	1.61	1.76	1.86	1.85	0.45	16.84

In the experiment, the significance of the two models for deriving the semantic knowledge in the minor semantic classes was also studied. It is a list of the minor semantic classes. Application of the semantic association model on them gave values > 0.3. Table 5 shows that the semantic association figures derived by the new model are not only much higher but also the new model can work out the semantic relationships of some classes which the simple co-occurrence model failed

to produce any values. The failure in the simple co-occurrence model is due to the fact that some semantic classes are only related indirectly e.g. Eb03 [密 疏] & Ed18 [細膩 粗疏].

Table 5 Minor Semantic Classes with Semantic Association Coefficient > 0.3

Semantic Class		Semantic Association Model	Simple Co-occurrence Model
Ke01 [喂 喳 喔 哼 噯]	Ke02 [嘿 嗯 嗨喲]	0.5148	0.3162
Ke02 [嘿 嗯 嗨喲]	Ke03 [哎呀 唉]	0.4954	0.1348
Ef02 [繁榮 蕭條]	Ih13 [新生 興起 衰落]	0.4327	0.3690
Ke01 [喂 喳 喔 哼 噯]	Ke03 [哎呀 唉]	0.4251	0.3090
He14 [足夠 相抵 剩餘 結存 虧欠]	If24 [賺 虧 損失 虧空]	0.4079	0.2974
Kd03 [啊 呢 喲]	Kd05 [嗎 的話 罷了 了 哉]	0.4065	0.2887
Kd04 [哩 唄 喲]	Ke02 [嘿 嗯 嗨喲]	0.3917	0.0845
Ba10 [它 甚麼]	Ed61 [這個 那個 某個 各個 其他 何]	0.3913	0.2500
Fa13 [放 堆 疊 掛]	Id07 [聳立 停放 懸掛 顛倒]	0.3709	0.2652
Ca06 [期限 期間]	Ca08 [中期]	0.3708	0.1091
Kb03 [跟 替 把 比]	Kb05 [為 被 以由]	0.3679	0.3131
Kd03 [啊 呢 喲]	Ke02 [嘿 嗯 嗨喲]	0.3670	0.0816
Eb12 [濃 稠 黏 稀]	Ec05 [深 淡]	0.3552	0.1715
Fc08 [呼 吐 呼喚 吹口哨]	Ke01 [喂 喳 喔 哼 噯]	0.3542	0.0838
Aa04 [他 他們]	Ba10 [它 甚麼]	0.3517	0.1690
Eb12 [濃 稠 黏 稀]	Ec11 [醇厚 油膩 清淡]	0.3506	0.2425
Kd03 [啊 呢 喲]	Kd04 [哩 唄 喲]	0.3506	0.2070
Eb09 [緊 鬆]	Ed34 [緊密 鬆散]	0.3405	0.2985
Ba09 [擔子 馱子 祭品]	Fa16 [包 裝 卷]	0.3363	0.3015
Kd04 [哩 唄 喲]	Ke01 [喂 喳 喔 哼 噯]	0.3361	0.0488
Kd03 [啊 呢 喲]	Ke01 [喂 喳 喔 哼 噯]	0.3350	0.0471
Ca06 [期限 期間]	Cb05 [內 外]	0.3332	0.2620
Ea04 [廣闊 寬敞 狹窄]	Ed38 [廣泛 無窮 狹隘]	0.3322	0.3021
Kb05 [為 被 以由]	Kc09 [因為 所以]	0.3321	0.2226
Ca06 [期限 期間]	Da05 [過程 內中]	0.3319	0.2994
Kd04 [哩 唄 喲]	Ke03 [哎呀 唉]	0.3234	0.1140
Kd03 [啊 呢 喲]	Ke03 [哎呀 唉]	0.3226	0.1101
Kd05 [嗎 的話 罷了 了 哉]	Ke02 [嘿 嗯 嗨喲]	0.3222	0.0707
Ea05 [深 淺 厚 薄 扁]	Ec05 [深 淡]	0.3220	0.1508
Ec05 [深 淡]	Ed17 [淡薄 膚淺 輕微]	0.3183	0.1231
Ab02 [老人 成年人 老小]	Ah02 [曾祖 祖父 祖母]	0.3173	0.1059
Ia08 [作響 爆炸]	Kf14 [嘟嘟 轟隆]	0.3118	0.0962
Ea03 [大 中 小]	Ed38 [廣泛 無窮 狹隘]	0.3107	0.0497
Ca06 [期限 期間]	Cb04 [前 後 中]	0.3106	0.1398
Eb03 [密 疏]	Ed18 [細膩 粗疏]	0.3102	0.0000
Fb04 [跨 越]	Je09 [支配 超越]	0.3086	0.1782
Cb27 [道路 路線]	Dg07 [途徑]	0.3074	0.2339
Kb03 [跟 替 把 比]	Kc01 [和 或者]	0.3074	0.2254
Bb03 [一堆 一把]	Fa16 [包 裝 卷]	0.3058	0.0909
Db02 [根源 來源]	Ja05 [起源 屬於]	0.3043	0.2722
Kd02 [來著 過]	Kd03 [啊 呢 喲]	0.3023	0.2108
Bb03 [一堆 一把]	Bc04 [蓋 桿 柄]	0.3022	0.1316

Figure 7 compares the two models by the relative frequency distribution of the semantic association and simple co-occurrence coefficients. Refer to the graph in Figure 7, the relative frequency of the semantic association coefficients is more elaborated than that of the co-occurrence coefficient. It can be observed that: (1) the former peaked at 9% and the latter at 0.5%, and (2) the interval of the frequently occurred coefficients were [0.005, 0.065] to [0.005, 0.1] for the simple co-occurrence and semantic association models, respectively. These imply that the coefficients in the semantic association model are generally larger than that in the simple co-occurrence model. From the above, it is clear that our new model can extract much more semantic knowledge from the CILIN than Lua' s model.

As mentioned before, the new model can measure weak as well as the strong semantic relationships. Therefore, the semantic association coefficients derived from it are more effective than the corresponding simple co-occurrence coefficients. This is reflected in Figure 8:

(1) For every coefficient in the co-occurrence model, there is a corresponding in the semantic association model.

(2) For the same semantic class, semantic association coefficient is always larger than or equal to the corresponding co-occurrence coefficient. This is clearly illustrated in the figure as all data points lie on or above the x=y line. In fact, we have measured the correlation between the two models. The correlation coefficient is 0.69.

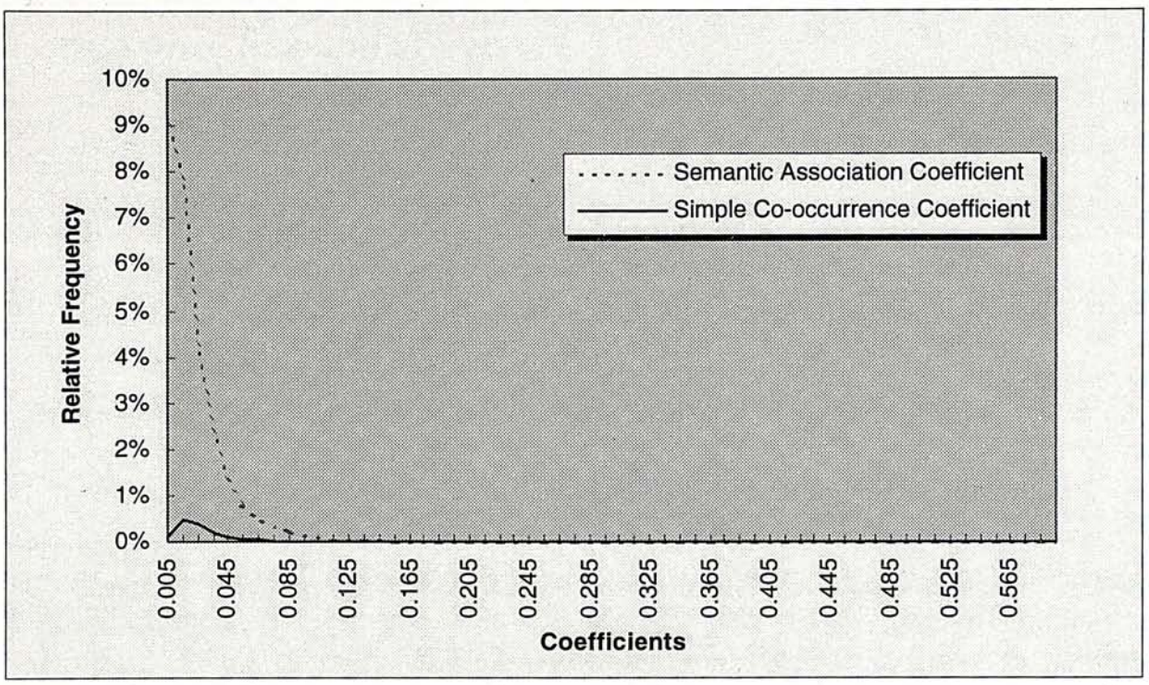


Figure 7 Graph Plotting the Relative Frequency Distribution of the Simple Co-occurrence Model and Semantic Association Model

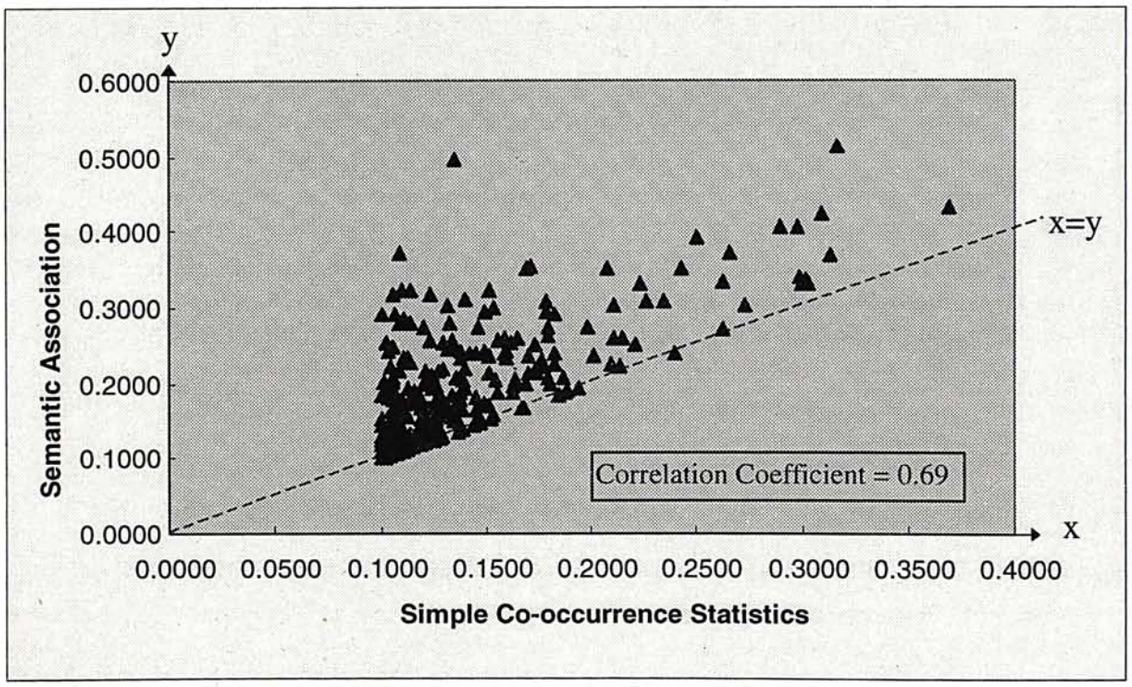


Figure 8 Graph Plotting Value of Semantic Association Value Determined by Semantic Association Model and Simple Co-occurrence Model (for the case of value > 0.1)

3.4 Inter-word Semantic Function

An *inter-word semantic function* is formulated. It is used to derive the inter-word semantic closeness (i.e. the connection weight, see Figure 3). For the formulation, the semantic classification provided by the CILIN and the association network constructed from the semantic association model described in the previous section are used. For a pair of words belonging to a single semantic class, their inter-word semantic relationship is equivalent to the semantic relationship between the semantic classes of these words. However, if the words in question belong to multiple semantic classes, the inter-word semantic relationship between these words can be quite complicated. For a pair of words with *N* and *M* semantic classes, there can be up to a total of *NxM* possible relationships between them. To measure the average relatedness between these words, a function is needed.

There are a number of conventional methods that could be used for computing the average relatedness e.g. simple average, geometrical average, etc. The desired function must satisfy two conditions.

(1) The function must give a value of 1 (i.e. max. relatedness) if the words are the same.

(2) In the case where the words have unique semantics, the function must return a value equivalent to the semantic relationship between their semantic classes.

As a result, the inter-word semantic function between words W_x and W_y, $Closeness(W_x, W_y)$, is defined as follows:

$$Closeness(W_x, W_y) = \frac{1}{|W_x| + |W_y|} \left(\begin{array}{l} \sum_{\forall SC_x \text{ of } W_x} MAX_{\text{for all } SC_y \text{of } Wy} Relatedness(SC_x, SC_y) \\ + \sum_{\forall SC_y \text{ of } W_y} MAX_{\text{for all } SC_x \text{of } Wx} Relatedness(SC_x, SC_y) \end{array} \right) \quad \textbf{(4)}$$

where

$|W_x|$ and $|W_y|$ are the number of semantic classes of words W_x and W_y respectively.

$Relatedness(SC_x, SC_y)$ is the relatedness function defined in Section 3.3.2 Equation (2).

This function satisfies the two necessary conditions and therefore is suitable for measuring inter-word semantic relationship. The proof is given as follows:

Case 1 - Two identical words

Proof

Assume a word W has n semantic class, thus

$$|W| = n$$

and

$$\sum_{\forall SC \text{ of } W} MAX_{\text{for all } SC \text{of } W} Relatedness(SC, SC) = n$$

Therefore,

$$Closeness(W, W) = \frac{1}{n+n}(n+n) = 1 \qquad \square$$

Case 2 - Two words belonging to unique semantic classes

Proof

Now,

$$|W_x| = |W_y| = 1$$

$$\sum_{\forall SC_x \text{ of } W_x} MAX_{\text{for all } SC_y \text{of } Wy} Relatedness(SC_x, SC_y) = Relatedness(SC_x, SC_y)$$

and

$$\sum_{\forall SC_y \text{ of } W_y} MAX_{\text{for all } SC_x \text{of } Wx} Relatedness(SC_x, SC_y) = Relatedness(SC_x, SC_y)$$

Therefore,

$$Closeness(W_x, W_y) = \frac{1}{1+1}\left(Relatedness(SC_x, SC_y) + Relatedness(SC_x, SC_y)\right)$$

$$= Relatedness(SC_x, SC_y) \qquad \square$$

Now, the inter-word semantic function can be used to calculate the inter-word semantic relationship between any pairs of word defined by the CILIN at run-time. A typical example of knowledge (i.e. the inter-word semantic relationship) acquired by the automatic procedure described in this chapter is given Figure 9.

Although the semantic relatedness between semantic classes is extracted using the semantic association model, the simple co-occurrence model could also be used to acquire the same semantic knowledge from the CILIN. This could be achieved by simply replacing the $Relatedness(SC_x, SC_y)$ function in Equation (4) by the associativeness coefficient *A* proposed by Lua K.T. (see Section 3.1.3). However, the significance of the inter-word semantic relationship would depend on

the effectiveness of the simple co-occurrence model which, as shown previously, would be very low.

In the next two chapters, the semantic association model is applied to the Noun-Verb-Noun compound word detection and the word-sense disambiguation problems.

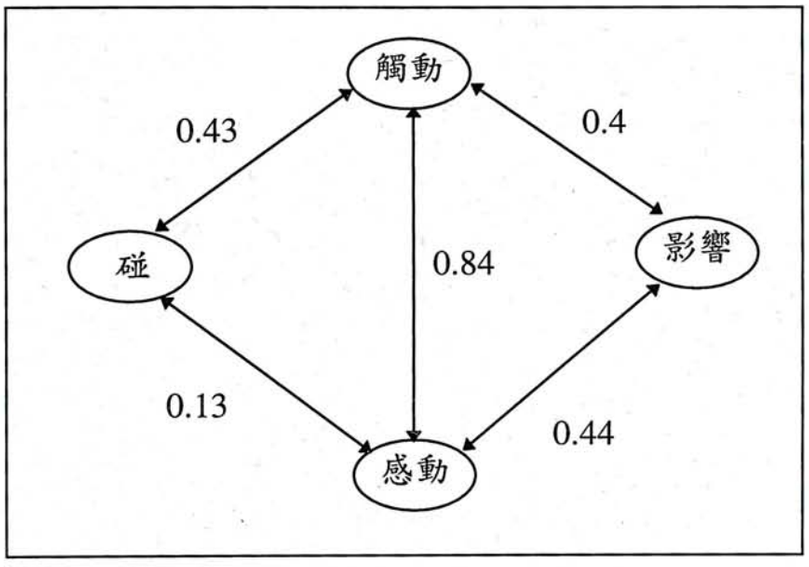


Figure 9 A Typical Inter-word Semantic Relationship Acquired from the CILIN

CHAPTER 4 NOUN-VERB-NOUN COMPOUND WORD DETECTION : AN EXPERIMENT

In this chapter, a specific NLP application - Noun-Verb-Noun compound word detection - is used as the testbed for comparing the semantic association model and the simple co-occurrence approaches. The major reason for selecting this application is that it extensively uses the inter-word semantic knowledge extracted from the CILIN. Consequently, the knowledge models defined by the two approaches can reflect from the effectiveness of the determination algorithm. Furthermore, N-V-N compound word detection is a hot issue in syntactical analysis of Chinese sentences.

Despite years of research, parsing of Chinese sentences still remains difficult. So far, no grammar rules that are general enough have been found for parsing open text. The reasons for this are:

(1) the lack of syntactic cues such as functional words and inflections, and

(2) the lack of grammar rules (from theoretical linguistic) in the Chinese language [Pun89].

The problem is escalated further as there exists numerous arbitrarily defined syntactical categories in Chinese, more than the same in other languages like English.

This is due to semantically driven nature of the Chinese nature language [Lua93b, Tang94].

To improve the situation, many researchers suggested to use algorithms to preprocess and identify local syntactic structures in a Chinese sentence, e.g. the method proposed by Pun and Lum [Pun89] which uses syntactic and semantic knowledge to parse noun phrases in a Chinese sentence. Similarly, the Noun-Verb-Noun (N-V-N) compound word is also one of the structures widely studied for the purpose of simplifying the syntactical analysis process.

In the next section, I will further elaborate and define the N-V-N compound word detection problem. And then, I will present a mathematical model which uses inter-word semantic relationship to identify the compound words. Later, in the final section of this chapter, I apply the model to a set of nouns and verbs and the effectiveness of the model is evaluated.

4.1 Overview

In English, a sequence of two or more words of certain categories can be combined together to form a structure that itself acts like a single word. For instance, "computer science" is a nominal compound (i.e. Noun-Noun compound) made up of two nouns. Compound words are frequently used in a sentence to enrich the semantic of the sentence. Similarly, in Chinese, a compound word is frequently formed by grouping together several words. For example, in the sentence:

我 最近 在 北京 開設 了 一 間 新 的 汽車 修理 公司。

(Recently, I opened a new car repairing company in Beijing.)

The 汽車修理公司 (car repairing company) is a Noun-Verb-Noun compound word. This compound is made up of three words: a Noun - 汽車(car), a Verb - 修理 (repairing) and another Noun - 公司(company). Based on the syntactical categories proposed by Tsinghua University [Tsinghua92], the sentence can be tagged as follows:

	1	2	3	4	5	6	7	8	9	10	11	12	13
	我	最近	在	北京	開設	了	一	間	新	的	汽車	修理	公司。
Part-of-speech tags:	r	t	p2ai vgn	s	vgn	nt	m	q	a	usde	ng	vgn	ng

Different from English, verbs and nouns in Chinese are sub-divided into sub-categories. The part-of-speech tags "vgn" and "ng" are examples of a sub-category of Verb and Noun, respectively. In the above example, within the 13 words, the word 在 belongs to multiple categories and, therefore, this example gives a total of 14 part-of-speech tags. If a new part-of-speech tag "n-v-n" is introduced to represent the N-V-N compound word (words 11, 12 and 13, the ng-vgn-ng combination), then tagging of the sentence will be simplified, viz.:

	1	2	3	4	5	6	7	8	9	10	11
	我	最近	在	北京	開設	了	一	間	新	的	汽車 修理 公司。
Part-of-speech tags:	r	t	p2ai vgn	s	vgn	nt	m	q	a	usde	n-v-n

With this new "n-v-n" tag, the total number of tags is reduced from 14 to 12. This, in turn, largely reduces the number of possible parsing trees. To further emphasize the advantage of N-V-N compound word detection, the above sentence is parsed using the *Dependency Grammar* proposed by Pan et al. [Pan91]. Dependency grammar has the following features in parsing [Section 2.1, Pan91] (see also Appendix A for more details):

- ✧ It uses a dependency tree to describe the inter-word functional relationship between all words in a sentence.
- ✧ Each link in a dependency tree directly describes the functional relationship between two words (e.g. "obj", "sub", and "nov" in Figure 11).
- ✧ The governor nodes on a dependency tree, which is generally a verb, indicates the head (i.e. the root of the tree).
- ✧ A dependency tree does not have any non-terminal nodes.

The dependency trees for the above example, with and without n-v-n, are given in Figure 10[7]. Using the dependency grammar, a n-v-n compound functions

[7] For simplication, the functional relationship between the words are omitted in Figure 10.

as a special Verb. Thus, the difference between the two final dependency trees are little. However, the number of intermediate trees varies greatly in the two cases. To simplify the analysis, the possible dependencies with the functional relationship given between the three "vgn[8]" are shown in Figure 11. From the figure, it is noted that the n-v-n compound reduces the number of possible dependencies by 60% (i.e. from 10 to 4). Therefore, if compound words can be identified early in parsing, syntax analysis of a sentence can be simpler and more efficient.

Although a pair of words can be grouped together when they are semantically coherent, human may never pair them up due to the usage of these words are unconventional. In other word, there are some other rules of thumb to determine the formation of compound words. As a result, the *usage* of a word can prohibit certain word combinations from occurring in practice. For instance, although a compound word "大學 修理 中心" (University Repairing Centre) is semantically acceptable in Chinese, this term will normally not be used in practice because, for it to mean a reparing centre within a university, the word "大學" is redundant and will be omitted usually. In view of this, formation of N-V-N compound words can be divided into three types: (a) both the semantic and usage are correct, (b) only the semantic is correct, and (c) neither the semantic nor usage is correct. Obviously, only types (a) and (b) are useful in NLP.

[8] Although the part-of-speech tag of the ambiguous word 在 is "p2as", it is known only after the evaluation of the parsing tree. Therefore, the "vgn" tag for this word is produced also in the intermediate stage.

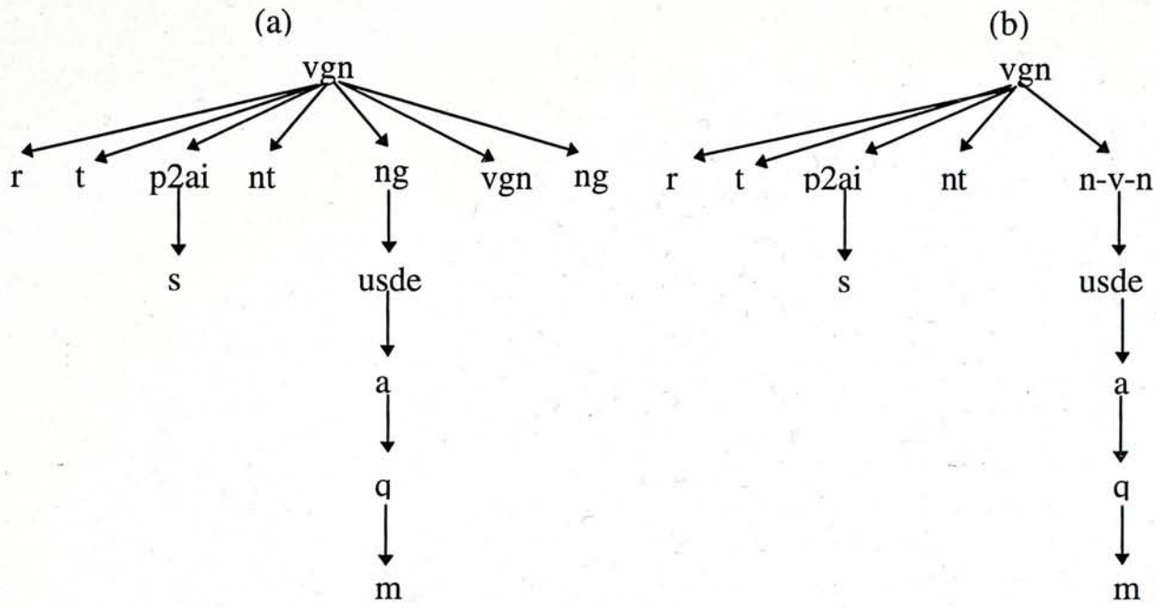


Figure 10 Dependency Trees for the Example (Functional Relationship Omitted)

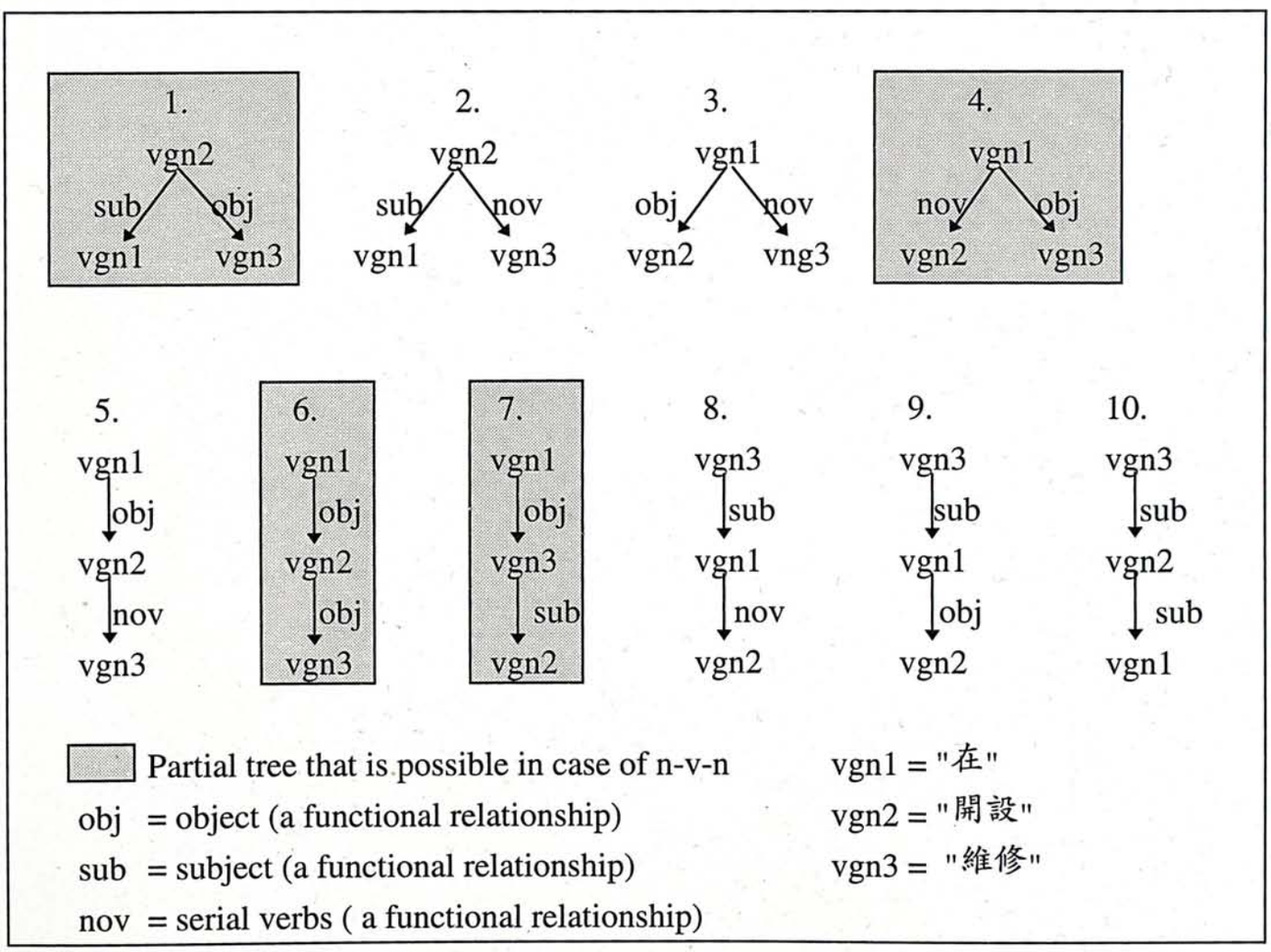
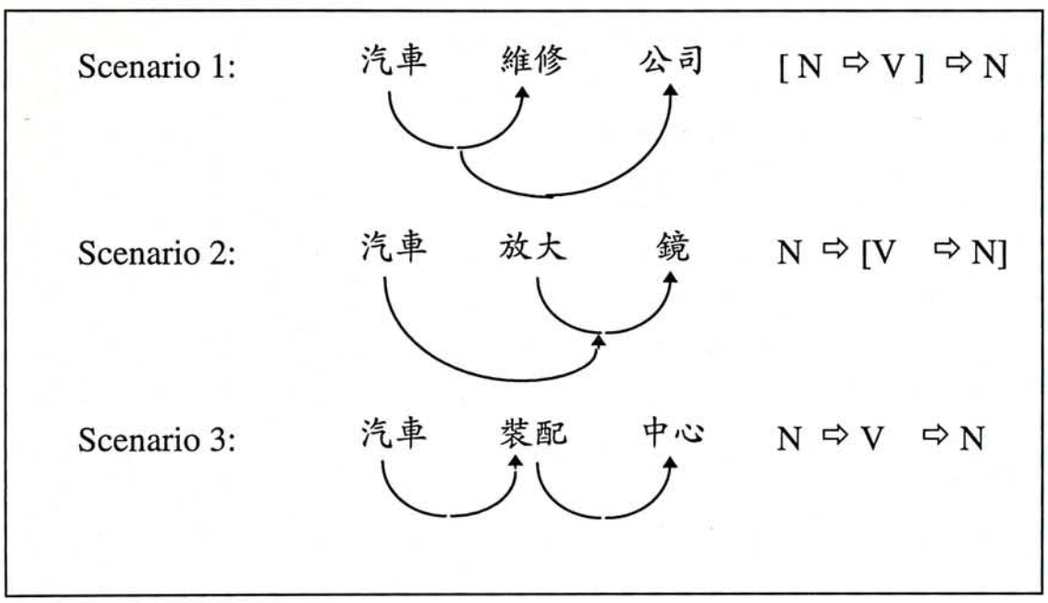


Figure 11 Intermediate Dependency Trees for the Example

4.2 N-V-N Compound Word Detection Model

Given a potential sequence of words, N-V-N compound word detection algorithm evaluates the possibility of these words forming a compound by analysing them in three different ways:



The double arrow represents relationship modifier (e.g. A⇨B means A modified B) and the square bracket implies word combination (e.g.. [A⇨B] means A + B). In the example, scenario 1 is read as 汽車 modifies 維修 and then [汽車+維修] modifies 公司. In scenario 2, 汽車 放大 鏡 is interpreted as 放大 modifies 鏡 followed by 汽車 modifies [放大+鏡]. In the last scenario, 汽車 modifies 裝配 and 裝配 modifies 中心 independently.

The N-V-N compound word detection uses the inter-word semantic knowledge extracted from the CILIN to evaluate the three scenarios. Note that a noun can modify a verb or a verb can modify a noun if and only if they have certain

inter-word semantic relationship (i.e. $Closeness(\ )>0$. See Equation (4) in Section 3.4). The mathematical formulation for each scenario is constructed as follows:

Scenario 1:

$$NVN_1(n_1, v, n_2) = \sqrt{Closeness(n_1, v) \times \frac{Closeness(n_1, n_2) + Closeness(v, n_2)}{2}}$$

Scenario 2:

$$NVN_2(n_1, v, n_2) = \sqrt{\frac{Closeness(n_1, v) + Closeness(n_1, n_2)}{2} \times Closeness(v, n_2)}$$

Scenario 3:

$$NVN_3(n_1, v, n_2) = \sqrt{Closeness(n_1, v) \times Closeness(v, n_2)}$$

where

n_1 is the first noun in the n-v-n sequence.

v is the verb in the n-v-n sequence.

n_2 is the second noun in the n-v-n sequence.

$Closeness(\ )$ is the inter-word semantic function defined in Section 3.4.

The actual score assigned to a potential N-V-N compound word sequence is computed by taking a simple average as given by the equation:

$$\boxed{NVN(n_1, v, n_2) = \frac{NVN_1(n_1, v, n_2) + NVN_2(n_1, v, n_2) + NVN_3(n_1, v, n_2)}{3}} \quad \textbf{(5)}$$

If the equation returns a value greater than 0, the words under consideration have potential to form a N-V-N compound word. Also, the higher the value is the more likely the words can form a compound.

4.3 Experimental Results of N-V-N Compound Word Detection

The primary objective of the experiment is to compare the semantic association model with the simple co-occurrence model. There are two measures that are frequently used to evaluate the performance of an information system model: *precision* and *recall*. Recall is the ratio of the correctly identified N-V-N compound words over all N-V-N compound words in the test set. Precision is the percentage of the identified compound words which are correct.

In the experiment, a total of 29 words were selected for testing (see Table 6). Equation (5) was used to identify n-v-n compound words. 10 of the 29 words are the first noun, 10 are verbs, and the remaining 9 are the second noun. A total of 900 (10 x 10 x 9) potential N-V-N compound words were generated and manually tagged. The tagged results were then classified into one of the following types: (a) both the semantic and usage are correct, (b) only the semantic is correct, and (c) neither the semantic nor usage is correct. The type (c) compound words were ignored. The statistics of the test set are given in Table 7.

The overall performance of the semantic association model and the simple co-occurrence model in terms of recall and precision is shown in Table 8. The results indicate that, for the first two types (i.e. a and b), the recall of the simple co-

occurrence model is between 2% to 3%. This is much smaller than 55% to 62% obtained by the semantic association model. On the other hand, the simple co-occurrence can obtain a higher precision (90% to 100%) than the semantic association model (51% to 78%) in the N-V-N compound word detection experiments. However, the difference in precision between the two models is inconclusive because the number of N-V-N detected by the simple co-occurrence is too small. Overall, the results can reinforce the conclusion drawn previously in Section 3.3.2 that the semantic association model can extract more semantic knowledge from the CILIN than the simple co-occurrence model. In fact, with a recall of 55% and a precision of 78%, the semantic association model performs rather well in N-V-N compound words detection.

Table 6 Words for Generating N-V-N Compound Words Testing Set

Candidate for First Noun (n_1)	Candidate for Verb (v)	Candidate for Second Noun (n_2)
火箭	修理	公司
桌子	維修	店
信息	裝配	中心
大學	發射	鋪
汽車	進口	員
機械	出口	場
電話機	運輸	器
玩具	放大	鏡
電腦	安裝	廠
電波	製造	

Table 7 Basic Statistics of the N-V-N Compound Word Testing Set

	Counting	Percentage %
Both Semantic and Usage are Correct	346	38%
Only Semantic is Correct	259	29%
Neither Semantic Nor Usage is Correct	295	33%
Total	900	100%

Table 8 Performance of the Semantic Association and Simple Co-occurrence in N-V-N Compound Word Detection

	Correct in Semantic and Usage		Correct in Semantic only	
	Recall (%)	Precision (%)	Recall (%)	Precision (%)
Simple Co-occurrence Model	3	90	2	100
Semantic Association Model	62	51	55	78

CHAPTER 5 WORD SENSE DISAMBIGUATION : AN APPLICATION

In the previous chapter, the N-V-N compound word detection problem was used as the testbed to examine the effectiveness of the semantic association model. A model was built on the inter-word semantic function for identifying potential N-V-N compound words based on a given sequence of words. The experimental results show that the network contains sufficient useful semantic knowledge for this task.

In this chapter, I will further demonstrate how the semantic association model can be used to handle a much general and harder problem in NLP, namely, *Word-Sense Disambiguation* [Bruce95]. From the practical point of view, many tasks in NLP require word-sense disambiguation, including Machine Translation (MT), Information Retrieval (IR), etc. A word with unknown meaning is a major impediment to the processing or understanding of a piece of text[9] in general. In MT, the senses[10] of every word in the input text must be precisely determined before the semantic of the entire text can be deduced. In IR, automatic keyword indexing, would be difficult if the keywords have multiple meanings.

[9] A text can be a phrase, a sentence, an article or a document

[10] "Sense" is a synonym of "meaning" and "semantic". These 3 words are used interchangeably in this thesis.

In the first section, I will define the word-sense ambiguity problem and then provide an overview of some contemporary word-sense disambiguation scheme. In the second section, I will provide a step-by-step discussion on how a linguistic-based model is constructed to assign a sense to an ambiguous word in Chinese full text using the semantic knowledge acquired from the CILIN. Finally, the proposed model is applied to determine the meaning of ambiguous words in newspaper articles and the results will be discussed in the last section.

5.1 Overview

What is word-sense ambiguity? An example can make it clear. For instance, consider the following two sentences:

1.	計算機 螢光屏 對 視力 也 有 不良 影響。 Computer screen also has adverse effect on vision.
2.	這些 都 是 影響 附會 之 談， 你 不 用 理會。 This is all rumour and hearsay and you need not to care about it.

The word *affect* (影響) is a homonym[11] (i.e. words with several distinct meaning). It has two senses, namely "the effect on other people or things" and "rumour; without evidence" for the above sentences, respectively. If the sense of *affect*is not differentiated in the IR systems (i.e. one index is used for both sentences), a query

[11] In general, words with two or more senses are referred to as polysemys and words that are the same in form or sound as another but different in meaning are referred to as homonyms. However, these are by no means standard definitions. These two terms are used interchangeably in this thesis to mean words having multiple senses.

containing this term will be unable to be processed precisely or accurately. This seriously affects the performance of the IR system and can in turn render it ineffective and useless for the application. Similarly, if the correct sense of *affect*is not identified in an MT application, this would result in erroneous translation.

The process of sense determination is often referred to as *word-sense disambiguation* [Bruce95]. In IR applications, a homonym would have multiple indexes - one for each of its sense. Word-sense disambiguation helps select one of the senses (i.e. indexes) which is most relevant to the context of the text. There are a number of prototypical word-sense disambiguation systems [Harder93, Wilks93, Niwa94]. Although these systems are different in implementation details, they share the same operational concept:

Consider a sentence[12] , *T*, which is composed of a sequence of n words:

$$T = W_1^{in} W_2^{in} W_3^{in} W_4^{in} W_5^{in} \ldots W_k^{in} \ldots W_n^{in}.$$

e.g. *T*=計算機 熒光屏 對 視力 也 有 不良 影響

Assume that W_k^{in} $\left(\text{e.g. } W_8^{in}, \text{ "affect"}\right)$ is a homonym which has the following senses - senses of a word is derived from a standard dictionary e.g. LDOCE:

$$W_k^{in} = S^{k,1} S^{k,2} \ldots S^{k,j} \ldots S^{k,m}$$

$$\text{e.g. } W_8^{in} = \text{ "affect" } = S^{8,1} S^{8,2} S^{8,3}$$

[12] For simplicity, but without loss of generality, the input text unit is a sentence in this discussion. In reality, it can well be a phrase, an article, a document, etc.

Further, each sense, say $S^{k,j}$, is itself a piece of text (i.e. it is composed of a sequence of p words), i.e. :

$$S^{k,j} = W_1^{k,j} W_2^{k,j} \ldots W_p^{k,j}$$

e.g. $S^{8,1} = W_1^{8,1} W_2^{8,1} W_3^{8,1} W_4^{8,1} W_5^{8,1} W_6^{8,1} W_7^{8,1} W_8^{8,1}$

= 對 別人 的 思想 或 行動 起 作用

e.g. $S^{8,2} = W_1^{8,2} W_2^{8,2} W_3^{8,2} W_4^{8,2} W_5^{8,2} W_6^{8,2} W_7^{8,2}$

= 對 人 或 事物 所起 的 作用

e.g. $S^{8,3} = W_1^{8,3} W_2^{8,3} W_3^{8,3} W_4^{8,3} W_5^{8,3}$

= 傳聞 的 ; 無 根據 的

The idea of word-sense disambiguation is to correlate the words in each $S^{k,j}$ (from j = 1 to m), in turn, with the words in T and to determine the similarity between the sense $S^{k,j}$ and the input sentence T. In general, the most similar sense will be chosen as the meaning of the word in question, i.e. W_k^{in}.

Existing word-sense disambiguation systems have different algorithm in determining the similarity between T and $S^{k,j}$. Similar to knowledge acquisition, these algorithms can be grouped into 2 categories: the linguistic-based approach, e.g. Niwa's algorithm using distance vectors [Niwa94]; and the corpus-base approach, e.g. Niwa's algorithm using co-occurrence vectors [Niwa94]. In practice, the linguistic-based approach is superior to the corpus-based approach in certain aspects. The

latter operates under the *majority win* principle. As a result, minority meanings may never be selected. On the other hand, the linguistic-based approach usually bases its analysis on standard linguistic resources (e.g. dictionary). Also, corpus are domain specific and as such the word-sense disambiguation system will not be able to handle any words which it has not *learned.* This is not so in the linguistic-based approach as the linguistic resources mostly contain universal (i.e. domain independent) information. Furthermore, in order to achieve high accuracy, the corpus must be sizable. Nevertheless, it is sometimes difficult to acquire a large set of domain specific text data. Even if this is possible, the time it takes to build up the co-occurrence statistics database can be very long.

Example of a linguistic-based sense disambiguation algorithm are [Harder93, Wilks93, Niwa94]. The Wilk's algorithm [Wilks93] uses *relatedness measure* to determine word similarity. The relatedness measure of 2 words was in fact the co-occurrence statistics within LDOCE, e.g. if "ball" and "sport" appeared in the same phrase 1000 times in the entire dictionary and "ball" and dancing" 20 times, the former word pair would have a higher relatedness measure than the latter. Experimental results showed that Wilks' algorithm had a 45% accuracy rate. Harder [Harder93] adopted a similar technique as Wilks by using co-occurrence statistics to measure word similarity. In addition, to increase the overall accuracy of disambiguation, syntactical information was used in Harder's algorithm. Experiments on 4 English verbs showed that the accuracy rate was 50%.

Different from the Wilks' and Harder's approaches, Niwa and Nitta [Niwa94] proposed to use *distance vector* to determine word similarity. A distance vector[13] measured the level of reference between two words. For example, in the following meaning:

> Library:- A collection of books for reading and borrowing.
>
> Book :- A series of written or printed or plain sheets of paper fastened together at one edge and enclosed in a cover.

the distance vector between "library" and "book" would be 1, between "library" and "paper" would be 2 and between "reading" and "paper" would also be 1. Since the definition of the word "library" contains the word "book", the distance between them is 1. Similar, the word "library" has the word "paper" in the definition and therefore is also 1. Although the definition of the word "library" does not contain the word "paper", its definition contains the word "book" and the "book" in turn has the word "paper" in its definition. So, the distance between the word "library" and the word "paper" is 2 (i.e. 1+1) since they are linked indirectly. Niwa and Nitta applied the distance vector approach to disambiguate the senses of 9 Japanese words. Further, Niwa also applied the co-occurrence statistics approach (i.e. corpus-based) to the same set of words. The experimental results indicated that the linguistic-based (distance vector) approach for determination of word similarity was a feasible solution and its accuracy was comparable to the corpus-based (co-occurrence statistics) approach - both approaches could achieve between 60% to

[13] Note that distance vectors are given in many Machine Readable Dictionaries (MRD).

100% accuracy. The results also showed that the co-occurrence statistics approach was more effective (i.e. higher accuracy) than the distance vector approach; but this finding is skeptical for the testing sample was small (9 homonyms only) and each homonym under test had only 2 senses. The experiment was not comprehensive and therefore the results cannot be a reliable performance indicator.

5.2 Word-Sense Disambiguation Model

In this study, a linguistic-based word-sense disambiguation algorithm for Chinese texts, referred to as LSD-C (Linguistic-based Sense Disambiguation for Chinese), is presented. Compared to the conventional linguistic-based algorithms, the LSD-C algorithm uses two standard linguistic resources, namely a thesaurus and a dictionary. The thesaurus is used to determine the similarity measure. Since the inter-word semantic structures in the thesaurus are well defined and widely accepted, the similarity measures derived from it is more reliable than the same in other systems, e.g. Niwa' s.

5.2.1 Linguistic Resource

The heart of a linguistic-based word-sense disambiguation algorithm is some form of linguistic resources. Depending on the nature of the resources, they can determine the words that the algorithm can handle as well as their senses (i.e. meanings), the semantic relationships between two or more words, the part-of-speech tags of the words, the grammatical rules etc. Many algorithms simply use a dictionary, e.g. [Niwa94], and any additional inter-word information are derived from

it *manually* (e.g. distance vectors). The LSD-C algorithm uses 2 different Chinese linguistic resources: namely, a standard Chinese dictionary 《現代漢語詞典》 [ShangWu84] and the CILIN. Both of them are popular and are widely used in many Chinese societies (e.g. Hong Kong). Based on them, the LSD-C will not be domain specific and will be applicable to any general Chinese text. Furthermore, the advantage of using a second resource, namely the thesaurus, is to formalize the inter-word semantic relationship.

《現代漢語詞典》 is a dictionary for daily use. It provides the readers the sense of 56,000 Chinese words. The sense of a word is expressed either by (a) its synonyms or antonyms, or (b) phrase or sentences describing the meaning of the word. In addition to senses, homonyms, polysemys[14] and part-of-speech tags may be provided for each word. The CILIN serves as the source for the inter-word semantic relationship for LSD-C (refer to Chapter 3).

5.2.2 The LSD-C Algorithm

The LSD-C algorithm accepts a sentence as input, and for each homonym in the sentences, identifies the sense most appropriate (i.e. semantically related) to the sentence. It is divided into the following steps:

Step 1: Word Segmentation.

Unlike English where words are formed by alphabets and two words are separated by a space or a punctuation mark, Chinese words are formed by one or

[14] A word with different related sense is a polysemous word.

more characters and there is no clear separation between them. This gives rise to the problem of how to identify the boundary of each word in a sentence. LSD-C uses the sequential Maximum Matching algorithm [Wong95] to segment the input sentence. On the average, it can achieve a accuracy rate of 95%. For example, the sentence: "計算機熒光屏對視力也有不良影響。" will be segmented in the following way:

✧ "計算機　熒光屏　對　視力　也　有　不良　影響。" (i.e. "Computer screen also has adverse effect on vision.")

Step 2: Eliminate stop words.

Similar to other languages, some Chinese words mainly play a syntactic role within the context of the sentence i.e. they do not have much semantic value. These are the stop words[15]. They do not contribute much to the word-sense disambiguation process; adversely, accounting for them would have increased its complexity. For example, in the sentences "計算機　熒光屏　對　視力　也　有　不良　影響。". 對, 也, and 有 are treated as stop words. Semantics of the sentence can be resolved by only using the other 5 words. There is no standard definition of a stop word in Chinese. The list of stop words defined for the experiment is given in Appendix D.

Step 3: Identify the homonym[16].

[15] For information retrieval application, stop words are poor candidates for indexing.

[16] In fact, words with multiple sense in a dictionary are more often referred to as polysemys.

For each word in the sentence, look up its senses from the《現代漢語詞典》. For example, 計算機 (computer), 熒光屏 (screen), and 視力 (vision) are words with a single sense; and 影響 (affect) has three, namely:

1. 對別人的思想或行動起作用 (has effect on others' thinking or action)
2. 對人或事物所起的作用 (the effect on other people or things)
3. 傳聞的;無根據的 (rumour; without evidence)

Step 4: Polish each sense of the homonym.

For each sense of a homonym, the sentence is word segmented and the stop words are eliminated i.e. apply steps 1 and 2 on each of the senses of the homonym. For instance, the last sense of the word 影響 (affect) will be segmented into:

✧ 傳聞; 根據 (rumour; evidence)

The words 的 and 無 are stop words. They are simply ignored.

Step 5: Calculate the similarity score.

To determine the similarity score, $score_{k,j}$ between the j-th sense, $S^{k,j}$ of the k-th word, W_k^{in}, in the input sentence, T (see Section 5.1), the following equation is used:

$$score_{x,j} = \frac{1}{N_{x,j}} \sum_{h \neq k}^{N_{in}} \sum_{i}^{N_{k,j}} Closeness\left(W_i^{k,j}, W_h^{in}\right) \times SF\left(W_k^{in}, W_h^{in}\right) \qquad \textbf{(6)}$$

where

k is the k-th word position in the input sentence; the corresponding word is W_k^{in}.

j is the j-th choice of a homonym.

$N_{k.j}$ is the number of words in the j-th sense of the k-th word of the input sentence.

N_{in} is the number of words in the input sentence.

$Closeness(W_1, W_2)$ is the inter-word semantic function defined in Section 3.4.

$W_i^{k,j}$ is the i-th word in the j-th sense of the k-th word in the input sentence.

W_h^{in} is the h-th word of the input sentence; similarly, W_k^{in} is the k-th word of the same.

$SF(W_1, W_2)$ is the significance factor for indicating the importance of W_2 in the sense disambiguation process of word W_1 to be discussed later.

Equation (6) consists of 4 parts (from right to left): (a) the significant factor function, $SF(\ )$; (b) the inter-word semantic function $Closeness(\ )$; (c) summation, $\sum\sum$; and (d) normalization, $\frac{1}{N_{k,j}}$.

(a) Significance Factor, $SF(\ )$

In a sentence, particularly a very long one, authors commonly divide a sentence into a number of text trunks, e.g. by using a comma, to indicate logical

separation in context. Therefore, words within the same trunk should be more effective in resolving the sense of a word within the same trunk. The significance factor takes this into consideration. It weighs the score in the following ways:

$$SF(W_1, W_2) = 1; \quad \text{if } W_1 \text{ and } W_2 \text{ are within the same trunk}$$
$$= w; \quad \text{otherwise}$$

where w is a constant between 0 and 1 to be determined experimentally

W_1 and W_2 are words.

(b) The inter-word semantic function, $Closeness(\ )$

The inter-word semantic function is to evaluate the semantic relationship between words based on the semantic knowledge of the CILIN (see Equation (4) in Section 3.4).

(c) Summation and (d) Normalization

To work out the final similarity score, the $Closeness(\ ) \times SF(\ )$ product of all possible combinations formed from pairing a word from the sense and a word from the input sentence are added together. The result of the summation would depend on the number of words used by the dictionary (i.e. $N_{k,j}$), in this case 《現代漢語詞典》, in describing the sense. To normalize this effect, the summation result is divided by the total number of words in the sense ($N_{k,j}$). An example will be given in the next section (Section 5.2.3) showing how the similarity score for the senses of the word 影響 (affect) are calculated.

Step 6: Sense selection.

The sense with the highest score is assigned to the homonym. In case the algorithm fails to assign any sense (i.e. all scores are 0), it is counted as incorrect. Similar to the N-V-N compound word detection (Chapter 4), I intentionally avoid using any other semantic information, e.g. the sense preference given by the dictionary, to assign a sense. Without using such semantic information, we guarantee that the algorithm uses only the semantic knowledge provided by the CILIN (i.e. *Closeness*( )), and therefore it can reflect the usefulness of the semantic relationships extracted by the semantic association model.

5.2.3 LSD-C in Action

To demonstrate how the LSD-C determines the sense of an ambiguous word in a sentence, the sense of the word 影響 (affect) is worked out step-by-step in the following input sentence:

計算機荧光屏對視力也有不良影響。

Step 1: Word Segmentation. After word segmentation, the input sentence becomes:

計算機　荧光屏　對　視力　也　有　不良　影響。

Step 2: Eliminate Stop Words. The third word 對 (to), the fifth word 也 (also) and the sixth word 有 (has) word are stop words and are eliminated. The input sentence after this stage becomes:

1	2	3	4	5
計算機	熒光屏	視力	不良	影響。

Step 3: Identify the Homonyms.

	1	2	3	4	5
	計算機	熒光屏	視力	不良	影響。
No. of sense:	1	1	1	1	3

Step 4: Polish Each Sense of the Homonyms. The homonym under investigation is the 5-th word 影響 and has 3 senses: They are:

(1) 對別人的思想或行動起作用

(2) 對人或事物所起的作用

(3) 傳聞的;無根據的

The above sense definitions are then processed by repeating steps 1 and 2 i.e. the sentences are word segmented and the stop words are eliminated. For the three senses of 影響, the following segmented senses are produced:

(1) 別人　思想　行動　起　作用

(2) 人　事物　起　作用

(3) 傳聞　根據

Step 5: Calculate the Similarity Score. Equation (6) is applied to compute the similarity score between the senses and the sentence based on the coefficient of inter-word semantic relationship, i.e. $Closeness(\ )$, derived from the CILIN and summarized in Table 9.

Table 9 Coefficients of Inter-word Semantic Relationship for the Example

		$W_1^{5,1}$	$W_2^{5,1}$	$W_3^{5,1}$	$W_4^{5,1}$ $W_3^{5,2}$	$W_5^{5,1}$ $W_4^{5,1}$	$W_1^{5,2}$	$W_2^{5,2}$	$W_1^{5,3}$	$W_2^{5,3}$
		別人	思想	行動	起	作用	人	事物	傳聞	根據
W_1^{in}	計算機	-	0.015	-	0.003	0.011	0.009	0.088	-	0.003
W_2^{in}	荧光屏	-	-	-	-	-	-	-	-	-
W_3^{in}	視力	-	0.039	0.012	0.010	0.067	0.045	0.038	0.009	0.016
W_4^{in}	不良	0.004	0.029	0.043	0.049	0.030	0.046	0.014	0.002	0.023

The similarity score for the first sense of 影響, i.e. $score_{5,1}$, is calculated as follows:

(a) Significance Factor

Since the sentence has only one trunk (i.e. the sentence consists of no sub-clauses), the significance factor $SF\left(W_k^{in}, W_h^{in}\right)$ is 1 for all words e.g. $SF\left(W_5^{in}, W_h^{in}\right) = 1 \quad \forall h \in \{1,2,3,4\}$

(b) The inter-word semantic function

Semantic closeness, $Closeness\left(W_i^{5,1}, W_h^{in}\right)$ $\left(\forall i \in \{1..5\} \text{ and } \forall h \in \{1..4\}\right)$, between the words of the polished sense and the words in the sentence extracted from Table 9 and are tabulated as follows:

		$W_1^{5,1}$	$W_2^{5,1}$	$W_3^{5,1}$	$W_4^{5,1}$	$W_5^{5,1}$
		別人	思想	行動	起	作用
W_1^{in}	計算機	-	0.015	-	0.003	0.011
W_2^{in}	荧光屏	-	-	-	-	-
W_3^{in}	視力	-	0.039	0.012	0.010	0.067
W_4^{in}	不良	0.004	0.029	0.043	0.049	0.030

(c) Summation

The $Closeness\left(W_i^{5,1}, W_h^{in}\right) \times SF\left(W_5^{in}, W_h^{in}\right)$ $\left(\forall i \in \{1..5\} \text{ and } \forall h \in \{1..4\}\right)$

product of the polished sense and the words in the sentence is obtained as follows:

	$W_1^{5,1}$	$W_2^{5,1}$	$W_3^{5,1}$	$W_4^{5,1}$	$W_5^{5,1}$
	別人	思想	行動	起	作用
W_1^{in} 計算機	-	0.015	-	0.003	0.011
W_2^{in} 荧光屏	-	-	-	-	-
W_3^{in} 視力	-	0.039	0.012	0.010	0.067
W_4^{in} 不良	0.004	0.029	0.043	0.049	0.030
$\sum_{i=1}^{5} Closeness\left(W_i^{5,1}, W_h^{in}\right) \times SF\left(W_5^{in}, W_h^{in}\right)$	0.004	0.083	0.055	0.062	0.108

Therefore,

$$\sum_{h \neq k}^{N_{in}} \sum_{i}^{N_{k,j}} Closeness\left(W_i^{k,j}, W_h^{in}\right) \times SF\left(W_k^{in}, W_h^{in}\right)$$

$$= \sum_{h \neq k}^{4} \sum_{i=1}^{5} Closeness\left(W_i^{5,1}, W_h^{in}\right) \times SF\left(W_5^{in}, W_h^{in}\right)$$

$$= 0.004 + 0.083 + 0.055 + 0.062 + 0.108$$

$$= 0.312$$

(d) Normalization

Since the polished sense consists of 5 words, $N_{k,j} = 5$. Hence, the similarity score for the first sense is:

$$score_{5,1} = \frac{1}{N_{k,j}} \sum_{h \neq k}^{N_{in}} \sum_{i}^{N_{k,j}} Closeness\left(W_i^{k,j}, W_h^{in}\right) \times SF\left(W_k^{in}, W_h^{in}\right)$$

$$= \frac{1}{5} \times 0.312$$

$$= 0.062 \quad \square$$

Similarly, the similarity scores for the other two senses are computed. The score of each sense and the ranking of the scores (in descending order) are given below:

Sense j	1	2	3
Similarity Score $Score_{5,j}$	0.062	0.102	0.026
Ranking	2	1	3

Step 6: Sense Selection. Since $Score_{5,2}$ is greater than $Score_{5,1}$ and $Score_{5,3}$, the second sense is selected as the final sense of 影響. In addition, the similarity scores also suggest that the third sense is the least likely choice for 影響 among the 3 given senses.

5.3 Experimental Results of Word Sense Disambiguation

Articles of various topics, including computer science, business, general studies, and etc., were selected from a Chinese newspaper for evaluating the LSD-C algorithm (see Appendix B). Instead of testing the algorithm with only selected words as in Niwa's experiment [Niwa94], all words with multiple meanings as defined in《現代漢語詞典》(refer to Appendix C) were disambiguated by the LSD-C algorithm. This gave a comprehensive evaluation of the domain coverage of the algorithm.

Basic statistics of the sample texts are tabulated in Table 10. The testing samples consist of six articles with a total of 54 sentences and 1439 words. A meaning of a homonym in the samples is defined by, on average, 3.95 words in 《現代漢語詞典》. For the 1439 words, 39.4% of them are stop words (i.e. these words have 0 sense, refer to Appendix D for the stop words used in the test), 34.68% of them have a single sense and the rest, 25.92%, have multiple senses, and 95% of the homonyms have less than 5 meanings (see Table 11). Dividing the number of characters of a word by its senses, it is interesting to know that on the average the highly ambiguous words (i.e. words with a large number of senses) are mostly one-syllabus characters. Detail statistics of the sentences are given from Table 12 to Table 17 and are further summarized in Table 18. The number of sentences in a sample ranges from 5 to 14. The percentage of homonym per sentence in the samples vary greatly from as low as 19.20% to as high as 35.23%. Based on the percentage, it is

noticed that sample 5 is the most while sample 3 is the least ambiguous articles in the testing pool. The average number of senses of a homonym is about 3.7.

Accuracy of the LSD-C algorithm in resolving the meaning of ambiguous words in each sample is given in Table 19 to Table 24. Table 25 summarizes the average performance of the LSD-C algorithm. The average accuracy of LSD-C algorithm ranges from 38.20% to 54.20% with an average of 48.00% depending on the degree of ambiguity of the text. Ranking the sample with the percentage of homonym[17] and the LSD-C accuracy (see Table 26) reveals that the accuracy of the LSD-C inversely related to the degree of ambiguity (i.e. the higher the homonym percentage is, the lower the LSD-C accuracy becomes). In other words, the LSD-C algorithm performs better for a less ambiguous text (i.e. lower % of homonym). This is what we expected.

[17] percentage of homonym = no. of homonym ÷ total no. of words in a document

Table 10 Summary Statistics of Testing Samples

Sample	Number			Average		
	Sentence	Characters	Words	Characters per Word	Senses per Word	Words per Sense
1	14	558	345	1.62	1.35	3.67
2	10	307	217	1.41	1.41	4.91
3	14	715	375	1.91	0.97	4.17
4	6	372	179	2.08	1.32	3.2
5	5	350	193	1.81	1.64	4.19
6	5	212	130	1.63	1.58	3.32
Overall	54	2514	1439	1.75	1.32	3.95

Table 11 Sense Statistics of Testing Samples

No. of senses	Frequency	Percentage (%)	Average no. of characters per sense
0*	567	39.40	1.45
1	499	34.68	2.25
2	153	10.63	1.9
3	60	4.17	1.7
4	66	4.59	1.21
5	21	1.46	1.14
6	25	1.74	1
7	24	1.67	1
8	13	0.90	1
9	3	0.21	1
> 10		0.56	1

* Stop words are assumed to be with zero sense.

Table 12 Sentence Statistics of Testing Sample 1

Sentence No.	No. of words	No. of homonym	% of homonym per sentence	Average no. of senses per word
1	44	15	34.09%	3.8
2	12	2	16.67%	3.5
3	8	1	12.50%	3
4	19	5	26.32%	3.2
5	13	5	38.46%	2.4
6	50	14	28.00%	3.5
7	11	4	36.36%	3.3
8	40	12	30.00%	3.7
9	25	6	24.00%	3.3
10	24	5	20.83%	4.4
11	11	2	18.18%	3.5
12	24	4	16.67%	2.5
13	14	1	7.14%	2
14	50	14	28.00%	4.8
Overall	345	90	26.09%	3.7

Table 13 Sentence Statistics of Testing Sample 2

Sentence No	No. of words	No. of homonym	% of homonym per sentence	Average no. of senses per word
1	21	9	42.86%	4.1
2	16	7	43.75%	3
3	37	11	29.73%	5.1
4	16	5	31.25%	4.2
5	30	6	20.00%	5.7
6	14	4	28.57%	4.8
7	21	7	33.33%	3.6
8	12	1	8.33%	10
9	26	5	19.23%	2.4
10	24	6	25.00%	5.7
Overall	217	61	28.11%	4.4

Table 14 Sentence Statistics of Testing Sample 3

Sentence No.	No. of words	No. of homonym	% of homonym per sentence	Average no. of senses per word
1	29	6	20.69%	3
2	33	4	12.12%	3
3	12	2	16.67%	3
4	28	6	21.43%	3.3
5	17	1	5.88%	3
6	35	9	25.71%	3.8
7	30	5	16.67%	2
8	20	6	30.00%	4
9	62	15	24.19%	4.3
10	23	0	0.00%	0
11	23	6	26.09%	2.2
12	15	3	20.00%	2.7
13	17	2	11.76%	3.5
14	31	7	22.58%	2.4
Overall	375	72	19.20%	3.3

Table 15 Sentence Statistics of Testing Sample 4

Sentence No.	No. of words	No. of homonym	% of homonym per sentence	Average no. of senses per word
1	28	3	10.71%	5.7
2	42	5	11.90%	3.2
3	39	12	30.77%	3.6
4	31	3	9.68%	3.3
5	22	9	40.91%	5.2
6	17	6	35.29%	4.8
Overall	179	38	21.23%	4.3

Table 16 Sentence Statistics of Testing Sample 5

Sentence No.	No. of words	No. of homonym	% of homonym per sentence	Average no. of senses per word
1	28	10	35.71%	3.9
2	61	26	42.62%	3.9
3	29	6	20.69%	3.2
4	27	6	22.22%	3.3
5	48	20	41.67%	3.6
Overall	193	68	35.23%	3.7

Table 17 Sentence Statistics of Testing Sample 6

Sentence No.	No. of words	No. of homonym	% of homonym per sentence	Average no. of senses per word
1	53	15	28.30%	4.2
2	26	12	46.15%	2.9
3	11	3	27.27%	2.7
4	13	4	30.77%	2.8
5	27	10	37.04%	3.1
Overall	130	44	33.85%	3.4

Table 18 Summary of Sentence Statistics of Testing Samples

Sample	No. of sentences	No. of words	No. of homonym	% of homonym per sentence	Average no. of senses per word
1	14	345	90	26.09%	3.7
2	10	217	61	28.11%	4.4
3	14	375	72	19.20%	3.3
4	6	179	38	21.23%	4.3
5	5	193	68	35.23%	3.7
6	5	130	44	33.85%	3.4
Overall	54	1439	373	25.92%	3.7

Table 19 Performance of the LSD-C Algorithm in Testing Sample 1

Sentence	No. of homonym	No. of correct	Percentage of correct %
1	15	7	46.70%
2	2	1	50.00%
3	1	1	100.00%
4	5	4	80.00%
5	5	2	40.00%
6	14	8	57.10%
7	4	2	50.00%
8	12	3	25.00%
9	6	5	83.30%
10	5	2	40.00%
11	2	1	50.00%
12	4	1	25.00%
13	1	1	100.00%
14	14	8	57.10%
Overall	90	46	51.10%

Table 20 Performance of the LSD-C Algorithm in Testing Sample 2

Sentence	No. of homonym	No. of correct	Percentage of correct %
1	9	4	44.40%
2	7	4	57.10%
3	11	9	81.80%
4	5	1	20.00%
5	6	0	0.00%
6	4	2	50.00%
7	7	2	28.60%
8	1	1	100.00%
9	5	2	40.00%
10	6	5	83.30%
Overall	61	30	49.20%

Table 21 Performance of the LSD-C Algorithm in Testing Sample 3

Sentence	No. of homonym	No. of correct	Percentage of correct %
1	6	3	50.00%
2	4	1	25.00%
3	2	1	50.00%
4	6	4	66.70%
5	1	0	0.00%
6	9	5	55.60%
7	5	5	100.00%
8	6	1	16.70%
9	15	10	66.70%
11	6	2	33.30%
12	3	1	33.30%
13	2	0	0.00%
14	7	6	85.70%
Overall	72	39	54.20%

Table 22 Performance of the LSD-C Algorithm in Testing Sample 4

Sentence	No. of homonym	No. of correct	Percentage of correct %
1	3	0	0.00%
2	5	2	40.00%
3	12	4	33.30%
4	3	3	100.00%
5	9	6	66.70%
6	6	3	50.00%
Overall	38	18	47.40%

Table 23 Performance of the LSD-C Algorithm in Testing Sample 5

Sentence	No. of homonym	No. of correct	Percentage of correct %
1	10	3	30.00%
2	26	7	26.90%
3	6	3	50.00%
4	6	2	33.30%
5	20	11	55.00%
Overall	68	26	38.20%

Table 24 Performance of the LSD-C Algorithm in Testing Sample 6

Sentence	No. of homonym	No. of correct	Percentage of correct %
1	15	6	40.00%
2	12	4	33.30%
3	3	2	66.70%
4	4	0	0.00%
5	10	8	80.00%
Overall	44	20	45.50%

Table 25 Overall Performance of the LSD-C Algorithm in Testing Samples

Sample	Sentence	No. of homonym	No. of correct	Percentage of correct %
1	14	90	46	51.10%
2	10	61	30	49.20%
3	14	72	39	54.20%
4	6	38	18	47.40%
5	6	68	26	38.20%
6	5	44	20	45.50%
Overall	54	373	179	48.00%

Table 26 Ranking the Sample by % of Homonym and LSD-C Performance

Sample	% of homonym		LSD-C performance	
	Value (from Table 18)	Ranking (in decreasing order)	Value (from Table 25)	Ranking (in increasing order)
1	26.09%	4	51.10%	5
2	28.11%	3	49.20%	4
3	19.20%	6	54.20%	6
4	21.23%	5	47.40%	3
5	35.23%	1	38.20%	1
6	33.85%	2	45.50%	2

CHAPTER 6 CONCLUSIONS & FURTHER RESEARCH

In this thesis, I have presented an approach to automatically acquire knowledge from a contemporary Chinese thesaurus 《同義詞詞林》(CILIN) for deriving the inter-word semantic relationship of Chinese words. In addition, I have shown in the previous chapters how to use the inter-word semantic relationship to tackle some well-known problems in Chinese NLP including the compound words identification and the word-sense disambiguation problems. Basic models for each of these problems were developed and tested. The first section of this chapter summarizes my achievements in this research study. I then identify and discuss in the second section some areas which are worth further investigation.

6.1 Conclusions

Linguistic knowledge is critical to the success of a NLP system. Although some efforts have been attempted to extract the knowledge from linguistic resources like dictionaries, thesauri, and encyclopedia, the achievements are far from being complete, particularly in the acquisition of knowledge for Chinese NLP. In this thesis, I propose a knowledge acquisition model to extract knowledge from the currently available standard Chinese natural language resources.

In Chinese, the CILIN 《同義詞詞林》 is a generally accepted Chinese thesaurus which contains valuable information about the word semantics. The major objective of this work is to develop a fully automatic knowledge acquisition technique to extract the knowledge about the inter-word semantic relationship from the CILIN. The major advantages of using knowledge acquisition techniques to obtain knowledge from dictionary and thesaurus include the following: (1) the technique is automatic (i.e. with little human intervention), and (2) the knowledge is comprehensive as it has a wide domain coverage. The main reason for using the CILIN as the knowledge source is simply because a lot of information regarding to semantic of words and inter-word semantic knowledge fundamental to many NLP applications are embedded in the CILIN. This information is at present not machine readable and cannot be applied to any NLP application.

The knowledge acquisition approach proposed in this study consists of two parts: (1) the design of an association network to represent the semantic knowledge, and (2) the design of an algorithm to extract the knowledge from the CILIN to construct the network and then to derive the inter-word semantic relationship from the network. The association network represents the semantic knowledge using a number of nodes of semantic classes inter-connected together with links with connection weights. Connection weights represent the association strength between semantic classes. A semantic association model is developed in the study to compute these connection weights. This model differs from the conventional simple co-occurrence approach in the ways it measures the association between semantic classes. Unlike the simple co-occurrence approach, the new model

derives the association between semantic classes based on their overlapping at semantic level rather than at symbolic level (i.e. words level). Both the *weak* and the *strong relationships* between semantic classes can be effectively measured from the CILIN under the new approach. The ability to identify weak relationship in the new method makes it superior than the simple co-occurrence approach.

To evaluate the significance of weak relationships, the Noun-Verb-Noun (N-V-N) compound word detection problem is used as a testbed. Currently, N-V-N compound word detection is one of the widely studied pre-processing techniques for simplifying syntactic analysis of Chinese sentences. A fundamental mathematical model has been established to identify the compound word for a given noun-verb-noun sequence. The mathematical model has been tested using the inter-word semantic relationship extracted from the CILIN by the semantic association model and the simple co-occurrence approach. Experimental results suggested that the weak relationship is equally important to the strong relationship. Without taking it into consideration, the *recall* for the N-V-N compound word experiment dropped dramatically from 55% to 2%.

To illustrate further how important the knowledge about the inter-word semantic relationship in NLP, I developed a model to tackle a major problem in NLP - the word-sense disambiguation problem. A linguistic based word-sense disambiguation algorithm, LSD-C, is proposed for resolving the meanings of an ambiguous word in a sentence. Currently, LSD-C is based only on the context of the sentence and the semantic knowledge provided by the CILIN for determining the

sense of the word defined in a standard Chinese dictionary 《現代漢語詞典》. Other additional linguistic clues such as syntactic categories are not used. Using LSD-C, an experiment has been carried out to disambiguate all the polysemys/homonyms occurred in six Chinese newspaper articles. In the experiment, LSD-C achieved an overall accuracy rate of 48% which is comparable to existing English word-sense disambiguation models.

6.2 Further Research

6.2.1 Enriching the Knowledge

In the semantic association model propounded in this thesis, semantic association between semantic classes is derived by performing one degree expansion on the semantic of the words in the semantic classes. This approach, in effect, can detect the indirect relationship (i.e. weak relationship) between semantic classes via one intermediate semantic class. If the relationship between the semantic classes was more circuitous i.e. it could be reached indirectly through more than more intermediate class, the present model would fail to acquire the semantic association. For instance, if a semantic class *A* is directly related to a semantic class *B* and the semantic class *B* is also directly related to another semantic class *C*, the model can show that the semantic class *A* is indirectly related to the semantic class *C* (i.e. *A-B-C*). However, if *C* is also directly related to *D* (i.e. *A-B-C-D*), the current model cannot detect the implicit relationship between *A* and *D*. This kind of weak relationships can extend further but the relationships between semantic classes becomes weaker and weaker as more intermediate semantic classes are involved. In fact, an approach called *spreading activation* [Cohen87] in information retrieval

works on a similar principle to identify conceptual related documents for a given query. It is hard to determine how many intermediate classes are passed through before a weak relationship becomes negligible. Nevertheless, its importance to NLP applications and its influence to the semantic association are worth further investigation.

The standard Chinese dictionary, 《現代漢語詞典》, and the CILIN were developed independently. In some cases, they are inconsistent; a word found in the dictionary may not exist in the thesaurus[18]. Since word segmentation, step 1 of the LSD-C algorithm, is based on the entries in the dictionary, some of the segmented words may have no semantic class. For example, the word 操作 (to operate) is undefined in the thesaurus. Furthermore, the standard Chinese dictionary, 《現代漢語詞典》 and the CILIN 《同義詞詞林》 were developed in 1983 and 1984, respectively. They cannot handle modern and technical terms like 鍵盤 (keyboard). To improve this situation, one can consider incorporating an additional domain specific dictionary to the system and at the same time enhance the thesaurus by adding more contemporary terms. However, enhancement to the thesaurus usually involves the work of lexicographer to reconstruct the association network. Unless new terms can be automatically and dynamically acquired and inserted into the network, the work of the lexicographer would be extremely complex and time consuming. At this time, there are still many open questions on where to acquire the knowledge and how to insert this into the network without any human involvement. These questions will remain open, at least, in the near future.

[18] Incidentally this is not a problem in English as MRD contains meanings of word as well as their semantic class information.

6.2.2 Enhancing the N-V-N Compound Word Detection Model

The primary objective of the N-V-N compound word detection model reported in this thesis is to compare the effectiveness of the semantic association model against the conventional simple co-occurrence model. It has been tested with an artificial set of word combinations (900 combinations) only. Although the recall and precision obtained from the experiment serve our own purpose, they are not enough to project the accuracy of the model in real cases. In reality, it is rare for some of the word combinations in the test set to occur in a sentence. For instance, it is meaningless in a sentence to have the N-V-N compound word "大學 出口器" . Had these meaningless words been identified and omitted prior to the application of our N-V-N compound word detection algorithm, the recall and precision rates would have been much higher. Nevertheless, at present, the model can already achieve 55 - 62% recall and 51 - 78% precision. These values are, in fact, practical.

The main source of error leading to the above is that some detected N-V-N compound words are correct in semantic but not in usage (see Section 4.1) Unfortunately, the CILIN provides only semantic knowledge but without word usage knowledge. One way to improve the precision of the semantic association model in detecting meaningless (i.e. unacceptable in usage) compound words is to enhance the model with knowledge about word usage. A corpus often contains the basic knowledge about word usage. Collocation statistics for N-V-N word sequence collected from the corpus could be combined with the present N-V-N compound word detection model to determine correct compound word usage.

To automate the extraction of collocation statistics, the N-V-N compound word detection model proposed can be used to bootstrap the knowledge acquisition procedure. It will pre-process the run text to construct a list of semantically feasible compound words. Invalid words (i.e. unacceptable in usage) can then be eliminated from the list.

6.2.3 Enhancing the LSD-C Algorithm

Similar to a dictionary in other languages, some homonyms are defined by very similar senses in the standard Chinese dictionary, 《現代漢語詞典》. For example, the word 作用 (effect) has the following four senses:

(1) Sense 1: 對事物產生影響

(2) Sense 2: 對事物產生某種影響的活動

(3) Sense 3: 對事物產生的影響; 效果; 效用

(4) Sense 4: 用意

In the above, the first three senses are very similar; they have the meaning of 影響 (affect). On the other hand, sense (4) is totally different from the others and means 用意 (intention). To differentiate among the first three senses, one needs to understand the semantic of each of them. Because this is a very complex task, most of the current word-sense disambiguation for English language ignores senses with similar meaning and only determine word senses using coarse grain meanings (i.e. refer to the above example, they treat the senses (1), (2) and (3) as the same meaning and the sense (4) as another). Therefore, a study can be done to exploit the potential of fine grain meanings (i.e. consider all the four senses separately). This should

lead to higher disambiguation accuracy.

In order to simplify computation, the LSD-C algorithm currently works under the basic assumption that the sentence under examination comprises only one ambiguous word. Other words in the same sentence all have unique sense. This assumption may still be acceptable if the ratio of homonyms to non-ambiguous words is low. However, other homonyms in the same sentence may severely affect the accuracy of the sense disambiguation algorithm if the ratio is high. Under this situation, a naive disambiguation algorithm which computes the closeness between the homonyms and all words in their corresponding senses[19] could lead to intolerably long. Therefore, the cost effectiveness of the present disambiguation algorithm to such situations should be carefully studied.

[19] This naive approach to determine once and for all the senses of all ambiguous words in a sentence is referred to as "Combinatorial Disambiguation".

APPENDICES

Appendix A - Dependency Grammar

Dependency grammar describes syntactic relationships or dependencies between pairs of words in a sentence. The structure of the parse tree produced by dependency grammar for a Chinese sentence is confined by the following fix axioms [Huang92b]:

- ✧ There is only one root in a sentence.
- ✧ Other elements (i.e. nodes of the dependency parse tree) must directly depend on another element of the same sentence.
- ✧ Any element strictly depends on only one other element.
- ✧ If the element A directly depends on an element B, and another element C is located between A and B in a sentence, then C must either be directly dependent of A or B, or an element which is between A and B.
- ✧ There is no direct dependence relationship between two elements which appear on the left and right branches of the root in a sentence.

Following the above axioms, each pair of words in a sentence will first be assigned with a governor and their dependence relationship is determined by their

syntactic function. For example, in the following sentence:

這 男孩 看見 那 老 人

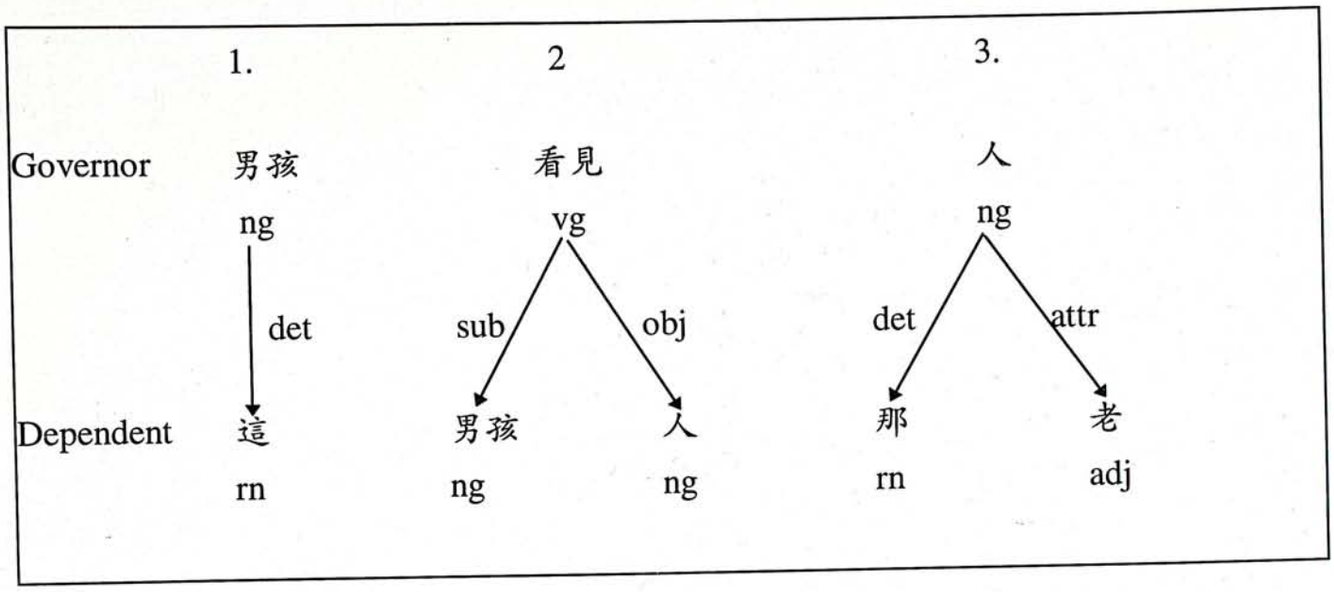
The boy sees the old man

Five governor-dependent pairs with four different functional relationships are produced, viz.:

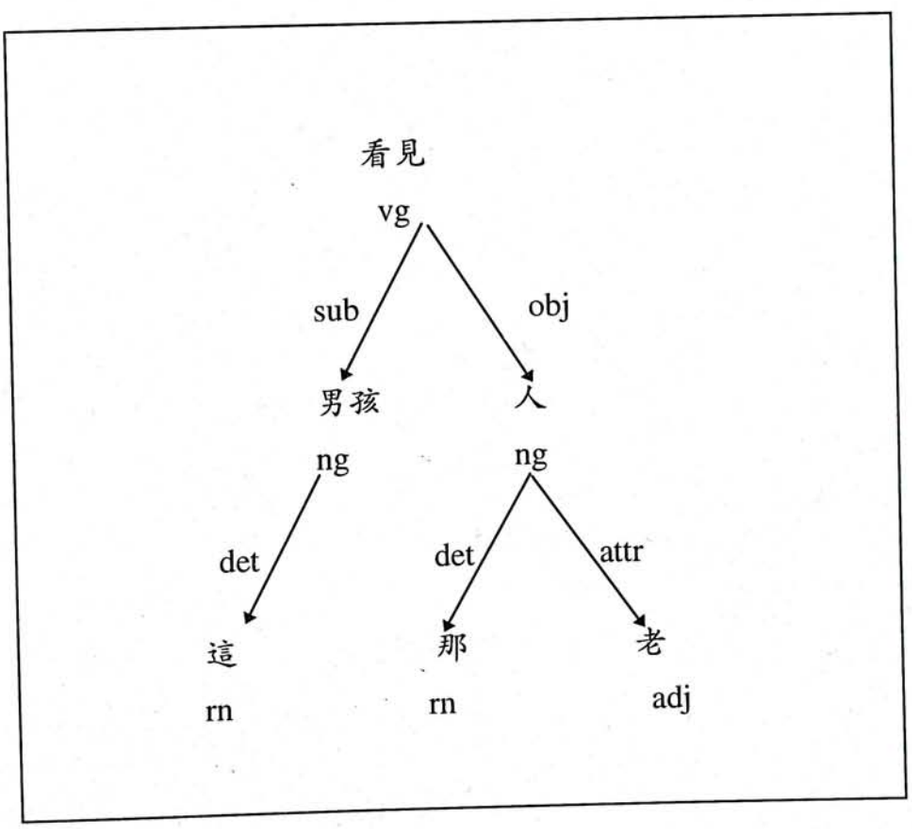
	1.	2	3.	4.	5.
Governor	男孩	看見	看見	人	人
	ng	vg	vg	ng	ng
	↓ det	↓ sub	↓ obj	↓ det	↓ attr
Dependent	這	男孩	人	那	老
	rn	ng	ng	rn	adj

where "ng", "vg", "rn", and "adj" are the part-of-speechs of the words, and "det", "sub", "obj", "attr" are the functional relationships between the words

Governor-dependent pairs with the same governor are then combined to form a smaller tree, viz.:



The combination processes continues until the final dependency tree is obtained, viz.:



Appendix B - Sample Articles from a Local Chinese Newspaper

🗁Sample 1

#1 當今人們坐在優雅舒適的辦公室裡，通過操作電子計算機、複印機和傳眞機等現代辦公設備處理業務，既方便工作效率又高，但若不注意，舒適的環境也會使人生病。

#2 長期操作計算機鍵盤，會導致手臂肌肉損傷和頸椎病。

#3 計算機熒光屏對視力也有不良影響。

#4 同時，許多計算機操作者還會出現頭痛、疲勞、易怒、失眠、心悸、厭食、惡心等症狀。

#5 據研究，這與電子計算機產生的輻射和電磁波污染有關 。

#6 由於複印機在工作時會產生大量臭氧，當臭氧濃度超過一定限度時，會嚴重危害人體的呼吸系統和神經系統，使操作者出現口腔和咽喉乾燥、咳嗽、胸悶和頭暈等症狀，甚至誘發氣管炎和中毒性肺氣腫等疾病。

#7 貼壁紙、鋪地毯，美化工作環境已成為時尚。

#8 然而， 在那些裝飾得豪華舒適的辦公室裡，從水泥、化纖織物、塑料製品和油漆塗料等建築裝飾材料中，會釋放出放射性氡、甲醛、苯系列等有害氣體，從而使人致病。

#9 氡是一種無色無味的氣體，它能通過呼吸道進入人體並沉積在肺組織內，破壞肺細胞，直到發生癌變。

#10 甲醛是一種溶劑，它會使人引起呼吸道不適，甚至導致哮喘和關節疼痛，而且甲醛也是一種致癌物質。

#11 而苯系列氣體則是眾所周知的致癌物。

#12 當然，人們在現代化的辦公室裡所出現的各種不適症狀，是由於多種因素的綜合作用而產生的。

#13 當我們了解這些危害後就應當積極地去預防它。

#14 比如，一次操作計算機和複印機的時不應過長，應注意工間休息和開窗換氣，有條件的話可購置換氣機、室內空氣淨化器等設備，改善室內空氣的質量，這些都是預防‘ 現代病 ’ 的有效措施。

🗁Sample 2

#1 患近視眼的人，如果不戴上眼鏡的話，是不易看清楚比較遠的東西的。

#2 他們看不清楚分明的輪廓，一切物體都只有模糊的外形。

#3 一個視力正常的人可以看清一棵大樹的細枝，患近視的人卻只到一片沒有顯明形狀的

綠色，細微的地方是完全看不到的。

#4 在夜裡，一切光亮的物體，對於患近視的人都變得很大。

#5 連夜裡的星空，他們看到的星星，不是幾千顆，只是幾百顆，而且他們看到的星卻比視力正常的人看到的要大。

#6 月亮在患近視的人看來顯得非常大而且非常近。

#7 這一切模糊、歪曲以及放大的原因，自然是由於患近視的人的眼睛構造上有毛病。

#8 他們的眼球曲率太大，也就是太凸出了。

#9 從外面物體上每一點所反射的光線通過眼球後，不能夠恰好集中在網膜上，而是在網膜的前面。

#10 因此光線射到眼球底部的網膜的時候，已經又散了開來，形成了模糊不清的像。

🗁Sample 3

#1 流傳在今年美國舉行的‘世界杯’足球比賽的參觀人士口中的最後的說話，恐怕將會是主辦國所採用的 CLIENT SERVER COMPUTING系統。

#2 預計全美將有三百六十萬名觀眾，以及一萬五千名流動的國際傳媒大軍，對於這種通常用於辦公室的電腦科技，這將是一次歷來最大型的考驗。

#3 該意念的心臟部分其實是一個相對簡單的概念。

#4 過去，大公司的電腦系統，傾向於包含大型的小電腦或立機電腦，並以一個相當呆板的形式供應笨拙的終端機網絡。

#5 有了 CLIENT SERVER SYSTEM，該等笨拙的終端機可由桌面私人電腦或工作站所代替。

#6 這些桌面電腦或工作站就是系統中的‘CLIENTS’，而服務它們的快捷的 SERVERS，是一些數據通訊儀器，它們的作用是迅速地散布資料和信息。

#7 桌面電腦純粹是一具標準的工業個人電腦或工作站，可採用日常的軟件或眾多特別為 CLIENT SERVER 網絡而特別設計的整套應用系統。

#8 因此新的系統應會更快捷、更平宜和比任何先前的系統更富於彈性。

#9 是次為美國主辦的‘世界杯’比賽貢獻出他們的電腦產品和知識的公司，包括 SUN MICROSYSTEMS（一間終端製造商﹏、EDS（一系統綜合公司﹏以及 SYBASE（一間數據庫發展公司﹏，他們聯合起來，在這項四年一度的球壇盛事進行期間，為從事球帶來一個複雜的、出息的電腦網絡。

#10 今年的‘世界杯’將會是歷來最複雜的一次，一個月內在九個不同的城市舉行五十二場比賽。

#11 主辦當局將須要追蹤超過三百五十萬張的入場券銷售，另外還要監督及運載超過二萬七千名志願工作人員往來比賽地點。

#12 主辦當局須要的是一個簡單、無須專門訓練也可使用的系統。

#13 所有比賽將會通過分配到九個城市去的一千台工作站及十五台 ＳＥＲＶＥＲＳ 播出。

#14 在每個比賽場地都有一部自學觸動式屏幕，讓記者可提取即時的資料，包括球員小傳、照片和統計，以至整隊球隊的資料等。

🗁Sample 4

#1 最近，北京城市數字數據網絡（ＤＤＮ）的第一階段工程已經完成，新網絡可為政府和金融部門提供一千二百個端口，服務於一千萬北京人。

#2 北京的這個合同是最大的兩個中國ＤＤＮ網合同之一；國際著名企業網絡公司－泰訊公司（ａｓｃｏｍ Ｔｉｍｅｐｌｅｘ）被指定為這兩個合同的供應商，一九九二年泰訊公司為中國的第一個高速局域ＤＤＮ網—上海郵電管理局提供設備。

#3 泰訊公司不僅贏得六個中國ＤＤＮ網絡合同，而且還成為國際衛星商業服務（ＩＢＳ）系統的Ｔ－１／Ｅ－１多路複用器獨家供貨商，這是一條源於中國的衛星數據專線服務主要通道。

#4 隨著中國金融部門和各大政府機構對於高速專線通信的需求增長，北京電信局迫切需要安裝高速的具有奶Ｑ七個節點的ＤＤＮ網絡。

#5 新網絡以２．０４８Ｍｂｐｓ的速率傳輸數據，比北京現有的模擬通信系統傳輸速率快二百倍。

#6 該網絡利用泰訊公司的孕ｘ ＬＩＮＫ／１００+ 作為主交換節點機，連接奶Ｑ三台ＬＩＮＫ／２+ 節點機。

🗁Sample 5

#1 中國最大規模的郵電部網管中心的數據採集，檢索與通訊工作和自動化管理軟件開發工作已於九它~年初開始。

#2 郵電部網管中心與香港科聯系統有限公司（簡稱科聯…簽署數據處理和全國網管系統工程的合約，價值約合五十斥Ｕ美元，該系統將在全國數據監控和福建數據監控系統中心運行，覆蓋全國郵電部系統，提供數據通訊和辦公室自動化資訊方面的服務。

#3 該系統採用ＨＰ９０００／８００小型機幼Ｍ ——包括Ｈ４０Ｘ１，Ｇ４０Ｘ１，Ｇ３０Ｘ２，另外還有工作站兩套，預期於九它~至九五年完成郵電通訊網絡工程的開發工作。

#4 系統將透過ＨＰ ＲＯＵＴＥＲ及其它網絡產品，進行全國郵電系統網管，屆時將大大促進中國郵電管理電子化和自動化的步伐。

#5 科聯還將協助郵電網管中心作技術後勤支持和軟件輔助開發及諮詢工作，除系統集成、硬件與網絡裝置，以及軟件應用外，科聯也在聯線數據庫發展領域上，提供顧問與程式支持服務。

🗁Sample 6

#1 為進一步推動中國房地產業向規範化方向發展，國家土地管理局中國地產諮詢中心與中國國際貿易促進委員會海南省分部於今年四月三日至六日海南‘椰子節’期間，在海口市聯合主辦九屆海南﹏中國房地產展銷洽談會。

#2 本屆展銷洽談椰子節交流房地產及城市建設信息，促成商品房供需各方直接見面洽談。

#3 同時溝通城市建設、舊城改造合作各方的聯繫。

#4 屆時將請有關部門高級官員介紹房地產的政策法規。

#5 國內外房地產界專家，將現場進行設計、評估、促銷、招標、交易等多項服務，金融機構也將到場開展信貸業務。

Appendix C - Ambiguous Words with the Senses Given by 《現代漢語詞典》

坐	#1	把臂部放在椅子凳子或其它物體上
	#2	乘,搭
	#3	背對著某一方向
	#4	把鍋,壺等放在爐火上
	#5	槍炮由于反作用而移動; 建築物由于基礎不穩固而下沉
	#6	瓜果等植物結實
	#7	指定罪
雅	#1	合符規范的
	#2	高尚的; 不粗俗的
	#3	西周朝廷上的樂歌
	#4	敬辭
	#5	交情
	#6	平素
	#7	很; 極
辦公室	#1	辦公的地方
	#2	辦理行政性事務的部門
複	#1	重複
	#2	繁複
印	#1	政府機關的圖章
	#2	印子
	#3	留下痕跡,特指文字或圖畫等留在紙上器物
	#4	符合
機	#1	機器
	#2	飛機
	#3	事情變化的樞紐; 有重要關係的環節
	#4	機會
	#5	生活機能
	#6	能迅速適應事物的變化的; 靈活
設備	#1	設置以備應用
	#2	進行某項工作或供應某種需要　需的成套建築或器物
處理	#1	安排解決
	#2	處理決定
	#3	特定方法工作產品進行加工

方便	#1	便利
	#2	適宜
	#3	婉辭,指有富餘的錢
工作	#1	從事髒力或腦力勞動
	#2	職業
	#3	業務;任務
效率	#1	機械、電器工作時,有用功在總所佔的百份比
	#2	單位時間完成的工作量
高	#1	從下向上距離大; 離地面遠
	#2	高度
	#3	在一般標準或平均程度之上
	#4	等級在上的
	#5	敬辭,稱別人的事物
	#6	酸根或化合物中比標準含一個氧原子
環境	#1	周圍的地方
	#2	周圍的情況和條件
使	#1	派遣; 支使
	#2	使用
	#3	讓; 叫; 致使
	#4	假如
	#5	奉使命辦事的人
人	#1	能制造工具並使用工具進行勞動的高等動物
	#2	每人; 一般人
	#3	指成年人
	#4	指某種人
	#5	別人
	#6	指人品質, 性格或名譽
	#7	指人的身體或意識
	#8	指人手, 人材
手	#1	人體上肢前端能拿東西的部份
	#2	拿著
	#3	小巧而便于拿的
	#4	親手
	#5	用于技能、本領
	#6	擅長某種技能的人或做某種事的人
損傷	#1	損害;傷害
	#2	損失

病	#1	生理或心理上發生的不正常的狀態
	#2	心病;私弊
	#3	缺點;錯誤
	#4	禍害;禍國
	#5	責備;不滿
影響	#1	對別人的思想或行動起作用
	#2	對人或事物所起的作用
	#3	傳聞的；無根據的
疲勞	#1	因體力或腦力消耗過多而需要休息
	#2	因運動過度或刺激過強，細胞，組織或器官的機能或反應能力減弱
	#3	因外力過強或作用時間過久而不能繼續起正常的反應
怒	#1	憤怒
	#2	氣勢很盛
心悸	#1	心臟跳動很利害
	#2	心里害怕
厭	#1	滿足
	#2	因過多而不喜歡
	#3	憎惡
食	#1	吃
	#2	專指吃飯
	#3	人吃的東西
	#4	一般動物吃的東西；飼料
	#5	供食用或調味用的
	#6	月球走到地球太陽之間遮蔽了月球時
據	#1	佔據
	#2	貸借;依靠
	#3	按照;依據
	#4	可以用做証明的事物
研究	#1	探求事物的眞相、物質、規律等
	#2	考慮或商討
輻射	#1	以中心向各個方向沿著直
	#2	熱的傳播方式的一種
污染	#1	使沾染上有害物質
	#2	空氣,土壤,水源等混入對生物有害或破壞環境衛生的物質的現象
工作	#1	從事髒力或腦力勞動
	#2	職業
	#3	業務;任務

大量	#1	數量多
	#2	氣量大,能容忍
超過	#1	由某物的後面趕到它的前面
	#2	高出
一定	#1	規定的; 確定的
	#2	固定不變; 然
	#3	表示堅決或確定; 定
	#4	特定的
	#5	相當的
咽喉	#1	咽頭和喉頭
	#2	比喻形勢險要的交通孔道
乾燥	#1	沒有水份或水份很少
	#2	枯燥,沒有趣味
胸	#1	軀幹的一部份在頸和腹之間
	#2	指心裏
悶	#1	氣壓低或空氣不流通而引起的不舒暢的感覺
	#2	使不透氣
	#3	不吭聲; 不聲張
	#4	聲音不響亮
	#5	在屋裏呆著, 不到外面去
	#6	心情不舒暢; 心煩
	#7	密閉; 不透氣
誘發	#1	誘導啓發
	#2	導致發生(多指疾病)
性	#1	性格
	#2	性質具性能物質含有成分產生
	#3	思想感情方面表現
	#4	有關生物生植生殖性欲性欲
	#5	性別
貼	#1	把薄片狀的東西粘在另一個東西上
	#2	緊挨
	#3	貼補
	#4	津貼
壁	#1	墻
	#2	某些物體上作用像圍墻的部份
	#3	像墻那樣直立的山石
	#4	壁壘

環境	#1	周圍的地方
	#2	周圍的情況和條件
裝飾	#1	在身體或物體的表面加些附屬的東西
	#2	裝飾品
豪華	#1	生活過份鋪張; 奢侈
	#2	富麗堂皇; 過份華麗
油漆	#1	泛指油類和漆類塗料
	#2	用油或漆塗抹
	#3	用礦物顏料和干性油樹脂等制成的塗料
建築	#1	造房子,修路,架橋等
	#2	建築物
裝飾	#1	在身體或物體的表面加些附屬的東西
	#2	裝飾品
材料	#1	可以直接造成成品的東西
	#2	提供著作內容的事物
	#3	可供參考的事實
	#4	比喻適于做某種事情的人材
釋放	#1	恢復被拘押者或服刑者的人身自由
	#2	把所含的物質或能量放出來
放射性	#1	某些元素自動把原子核中的物質放射出來
	#2	醫學上指由一個痛點向周圍擴散的現象
致	#1	給與; 向對方表示
	#2	集中予某個方面
	#3	招致
	#4	以致
色	#1	顏色
	#2	臉上表現的神氣樣子
	#3	情景; 景象
	#4	物品的質量
	#5	指婦女美貌
無味	#1	沒有滋味
	#2	沒有趣味
通過	#1	從一端或一側到另一側
	#2	議案等經過法定人數的同意而成立
	#3	從一端或一側到另一端或另一側; 穿過
	#4	議案等經過法定人數的同意而成立
沉積	#1	河流流速減慢時,水中所挟帶的岩石、砂礫、泥土等沉淀下來,淤積河床

和海灣等低洼地帶

#2 指物質在溶液沉淀積聚的現象

組織 #1 安排分散的人或事物使具有一定的系統性或整體性

#2 系統; 配合關係

#3 紡織品經緯紗線的結構

#4 機體中構成器官的單位

#5 按照一定宗旨和系統建立起來的集體

破坏 #1 使建築物等損壞

#2 使事物受到損害

#3 變革

#4 違反

#5 損壞

發生 #1 原來沒有的事出現了; 產生

#2 卵子受精後逐漸生長的過程

變 #1 和原來不同

#2 改變; 變成

#3 使改變

#4 能變化的; 已變化的

#5 變實

#6 有重大影響的突然

#7 指變文

關節 #1 骨頭互相連接的地方

#2 起關鍵性作用的環節

#3 指舊時暗中行賄勾通官吏的事

致 #1 給與; 向對方表示

#2 集中予某個方面

#3 招致

#4 以致

物質 #1 獨立存在予人的意識之外的客觀實在

#2 特指金錢, 生活資料等

物 #1 東西

#2 指自己以外的人或跟自己相對的環境

#3 內容; 實質

當然 #1 應當這樣

#2 合於事理或情理, 沒有疑問

因素 #1 構成事物本質的成分

#2 決定事物成敗的原因或條件

綜合	#1	把分析的對象或現象的各個部份,各屬性聯合或統一的整體
	#2	不同種類,不同性質的事物在一起
作用	#1	對事物產生影響
	#2	對事物產生某種影響的活動
	#3	對事物產生的影響; 效果; 效用
	#4	用意
當	#1	相稱
	#2	應當
	#3	面對著; 向著
	#4	正在
	#5	擔任
	#6	承當
	#7	掌管
了解	#1	知道得清楚
	#2	打聽
積極	#1	肯定的; 正面的
	#2	進取的; 熱心的
去	#1	離開
	#2	失去
	#3	除去
	#4	距離
	#5	過去的
	#6	從所在地別的地方
工	#1	工人和工人階級
	#2	工作; 生產勞動
	#3	工程
	#4	工業
	#5	一個工人或農民勞動日的工作
	#6	技術和技術修養
	#7	長於; 善於
	#8	精巧; 精致
間	#1	中間
	#2	一定的空間或時間裡
	#3	一間屋子; 房間
	#4	空隙
	#5	隔開
	#6	挑拔使人不和; 離間

	#7 拔去或鋤去
開	#1 使關閉著的東西再不再關閉
	#2 打通; 開辟
	#3 舒張; 分離
	#4 解凍
	#5 解除
	#6 發動或操縱
	#7 開拔
	#8 開辦
	#9 開始
	#10 舉行
	#11 寫出
	#12 支付
	#13 開革; 開除
	#14 受熱而沸騰
	#15 喫
	#16 指十分之幾的比例
	#17 印刷上指相當于整張紙的若干分之一
換	#1 給人東西時從他那裡取別的東西
	#2 變換; 更換
	#3 兑換
氣	#1 氣體
	#2 特指空氣
	#3 氣息
	#4 指自然界冷熱陰晴等現象
	#5 味兒
	#6 人的精神狀態
	#7 人的作風習氣
	#8 生氣; 發怒
	#9 使人生氣
	#10 欺負; 欺壓
	#11 中醫指人體内器官正常地發揮機能的原動力
	#12 中醫指某種病象
條件	#1 影響事物發生; 存在或發展的因素
	#2 為某事而提要的求定或定出的標準
	#3 狀況
室	#1 屋子

	#2	機關;公社、工廠、學校等內部的工作單位
	#3	二十八宿之一
空氣	#1	構成地球周圍大氣的氣體
	#2	氣氛
器	#1	器具
	#2	器官
	#3	度量
	#4	器重
設備	#1	設置以備應用
	#2	進行某項工作或供應某種需要　需的成套建築或器物
質量	#1	物體中所含物質的量,也是物體慣性的大小
	#2	產品或工作的優劣程度產生

Appendix D - List of Stop Words for the Testing Samples

Although there is not a well defined stop word standard for word-sense disambiguation applications, it is well understood that stop words are the words that do not provide much contextual information for disambiguating neighbouring words. Based on this principle, a list of stop words, specific to the testing samples, was defined in the following for the word-sense disambiguation experiment.

２・０４８	大	任何	的話	最近
ＤＤＮ	小	先前	者	最後
ＥＤＳ	已	全	初	就
Ｇ３０Ｘ２	已經	再	近	就是
Ｇ４０Ｘ１	才	同時	非常	幾千顆
Ｈ４０Ｘ１	不	各	前	幾天
ＨＰ　ＲＯ	中	名	前面	幾百顆
ＨＰ９００	之	因	則	然
ＩＢＳ	之一	因此	卻	然而
ＬＩＮＫ／	五十二場	因而	很	等
ＬＩＮＫ／	五十四萬	地	後	著
Ｔ－１／Ｅ	五點一二	在	既	量
一	仍然	好	是	項
一九九二年	內	如	是次	須要
一千二百個	內外	如果	為	會
一千台	六日	成為	為例	當
一千萬	六個	有	甚至	當今
一片	及	此	皆是	當然
一次	反而	而	相反	經過
一些	太	而且	約	裡
一具	比	至	若	該
一度	比如	但	要	過去
一個	乏	即時	個	像
一隻	以	均	原因	對
一條	以上	我們	特別	對於
一部	以及	更	能	種
一棵	以至	每	能夠	與
一間	他們	沒有	高	需
一萬五千名	包含	那些	偏	需求
一輪	包括	並	將	需要
一點	去	依然	將於	應
九五年	可	例	得	應當
九四	可以	來	從	還
九四年	另外	兩個	從而	還有
九個	只	兩套	現	還要
了	只是	具	現今	簡稱
二百倍	四十三台	其他	現有	讓
二萬七千名	四月三日	其它	現時	無

十五台
十分
又
三九五
三九零
三八八點五
三八四
三百五十萬
三百六十萬
三種
上
也
也是
四台
四年
四套
外
它
它們
本
未
由
由此
由於
目前
亦
其實
到
和
居
居時
底
或
所
所有
於
於是
易
的
第一
第一個
細
被
許多
這
這些
這是
這個
部
都
最
最大

REFERENCES

[Agosti92] Agosti M. and Marchetti P.G., *User Navigation in the IRS Conceptual Structure through a Semantic Association Function*, The Computer Journal, Vol. 35, No. 3, pp. 194-199, 1992.

[Allen87] Allen J., *Natural Language Understanding*, The Benjamin/Cummings Publishing Company, Inc., 1987.

[Amsler89] Amsler R.A., *Words and Worlds*, Theoretical Issues in Natural Language Processing, Lawrence Erlbaum Associates, Publishers, Yorick Wilks Editor, pp. 11-15, 1989.

[Bates93] Bates M., Bobrow R.J., and Weischedel R.M., *Critical Challenges for Natural Language Processing*, Challenges in Natural Language Processing, Bates, Madeleine and Weischedel, Ralph M. (editors), Cambridge University Press, pp. 3-34, 1993.

[Berwick89] Berwick R.C., *Learning Word Meanings from Examples*, Semantics Structures, Waltz D. (editor), Lawrence Erlbaum Associates, Inc., pp. 89-101, 1989.

[Binot93] Binot J.L. and Jensen K., *A Semantic Expert Using an On-line Standard Dictionary*, Natural Language Processing: The PLNLP Approach, Kluwer Academic Publishers, pp. 119-147, 1993.

[Boguraev89a] Boguraev B.K., *On-line Lexical Resources for Natural Language Processing*, Recent Developments and Applications of Natural Language Processing, GP Publishing Inc., pp. 192-213, 1989.

[Boguraev89b] Boguraev B.K., and Briscoe T., *Computational Lexicography for National Language Processing: Introduction*, Longman, pp. 1-39, 1989.

[Bruce95] Bruce R., *A Statistical Method for Word-Sense Disambiguation*, Ph.D. Dissertation, New Mexico State University, 1995.

[Calzolari90] Calzolari N. and Zampolli A., *Methods and Tools for Lexical Acquisition*, Proceeding of EAIA' 90 2nd Advanced School on Artificial Intelligence, October 1990.

[Chodorow85] Chodorow et al., *Extracting Semantic Hierarchies from a Large On-line Dictionary*, Proceedings of the 23rd Annual Meeting of the ACL, pp. 299-304, 1985.

[Chomsky72] Chomsky N., *Language and Mind*, Enlarged ed. New York: Harcourt Brace Jovanovich, 1972.

[Chen94] Chen K. J., *Linguistic Information and Lexical Data Management in Electronic Dictionary Research*, Tutorial Notes, Proceedings of the ICCPOL' 94, Chinese Language Computer Society, pp. 22-28, 1994.

[Church89] Church K. W. and Hanks P., *Word Association Nouns, Mutual Information and Lexicography*, Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, pp. 76-83, 1989.

[Cobuild87] Cobuild, *Collins COBUILD English Language Dictionary*, William Collins Sons & Co. Ltd., 1987.

[Cohen87] Cohen P.R. and Kjeldsen R., *Information Retrieval by Constrained Spreading Activation in Semantic Networks*, Information Processing & Management, Vol. 23, No. 4, pp. 255-268, 1987.

[Crouch88] Crouch C.J., *A Cluster-Based Approach to Thesaurus Construction*, Proceedings of the SIGIR' 88, pp. 309-321, 1988.

[Crouch92] Crouch C.J. and Yang B., *Experiments in Automatic Statistical Thesaurus Construction*, Proceedings of the SIGIR' 92, pp. 77-88, 1992.

[Cullingford86] Cullingford R. E., *Natural Language Processing*, Rowman & Littlefield, 1986.

[DeJong89] DeJong G., *An Overview of the FRUMP System*, Strategies for NLP, pp. 149-175, 1989.

[Dik78] Dik S.C., *Stepwise Lexical Decomposition*, Reter de Ridder Press, Lisse, 1978.

[Fromkin93] Fromkin V. and Rodman R., *An Introduction to Language*, 5th Edition, Harcourt Brace Jovanovich College Publishers, 1993.

[GB92] GB 中華人民共和國國家標準, 《信息處理用現代漢語分詞規范》 (Contemporary Chinese Language Word Segmentation Specification), 國家技術監督局, 1992.

[Grishman86] Grishman R., *Computational Linguistics: An Introduction*, Cambridge University Press, 1986.

[Harder93] Harder L.B., *Sense Disambiguation Using On-line Dictionaries*, Natural Language Processing: the PLNLP Approach, Kluwer Academic Publishers, pp. 247-261, 1993.

[Huang92a] Huang C.N, Yuan C.F., and Pan S.M 黃昌寧, 苑春發及潘詩梅，《語料庫、知識獲取和句法分析》，中文信息學報, Vol. 6, pp. 1-6, 1992.

[Huang92b] Huang C.N., Wu S., and Yuan C.F., *A Study in the Robustness of Chinese Parser with the Variety of Knowledge*, International Symposium on Natural Language Understanding and AI, pp. 54-57, 1992.

[Kimoto90] Kimoto H. and Iwadera T., *Construction of a Dynamic Thesaurus and Its Use for Associated Information Retrieval*, Proceedings of SIGIR` 90, pp. 227-239, 1990.

[Klavans93] Klavans J., Chodorow M., and Wacholder N., *Building a Knowledge Base from Parsed Definitions*, Natural Language Processing: The PLNLP Approach, Kluwer Academic Publishers, pp. 119-147, 1993.

[Jensen93] Karen Jensen, *PEG: The PLNLP English Grammar*, Natural Language Processing: The PLNLP Approach, Kluwer Academic Publishers, pp. 29-45, 1993.

[Jacobs94] Jacobs P. S., *Word Sense Acquisition for Multilingual Text Interpretation* Proceedings of COLING' 94, pp. 665-671, 1994.

[Lenat86] Lenat, D., Prakash, M., and Shepherd, M., *CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks*, AI Magazine, 6(4), pp. 65-85, 1986

[Lu94] Lu C. and Yi C.H., *A Comprehensive Chinese Thesaurus System and its Weighting Scheme*, Proceedings of the ICCPOL` 94, Chinese Language Computer Society, pp. 289-294, 1994.

[Lua93a] Lua K.T., *A Study of Chinese Word Semantics*, Computer Processing of Chinese & Oriental Languages, Vol. 7, No. 1, June 1993, pp. 37-60, 1993.

[Lua93b] Lua K.T., 《漢字的聯想與漢語語義場》, Communications of COLIPS, Vol. 3, No. 1, pp. 11-30, 1993.

[Lua93c] Lua K.T., *A Study of Chinese Word Semantics and its Prediction*, Computer Processing of Chinese & Oriental Languages, Vol. 7, No. 2, December 1993, pp. 167-189, 1993.

[Mei83] Mei et al. 梅家駒，竺一鳴，高蘊琦，殷鴻翔, 《同義詞詞林》, 上海辭書出版社, 上海, 1983.

[Niwa94] Niwa Y. and Nitta Y., *Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries*, Proceedings of COLING' 94, pp. 304-309, 1994.

[Paice91] Paice C.D., *A Thesaurus Model of Information Retrieval*, Information Processing & Management, Vol. 27, No. 5, pp. 433-447, 1991.

[Pan91] Pan S.M. Yuan C.F., and Huang C.N., *Knowledge Acquisition from Analyzed Corpus for Chinese Parsing*, Natural Language Processing Pacific Rim Symposium, pp. 358-365, 1991.

[Patterson90] Patterson D.W., *Introduction to Artificial Intelligence and Expert System*, Prentice-Hall International Editions, pp. 264-266, 1990.

[Pun89] Pun K.H. and Lum B., *Resolving Ambiguities of Complex Noun Phrases in a Chinese Sentence by Case Grammar*, Computer Processing of Chinese & Oriental Languages, Vol. 4, Nos. 2 &3, July 1989, pp. 185-202, 1989.

[Pustejovsky91] Pustejovsky J., *The generative lexicon*, Computational Linguistics, Volume 17, Number 4, pp. 409-441, 1991.

[Pustejovsky93] Pustejovsky J., Bergler S., Anick P., *Lexcical Semantic Techniques for Corpus Analysis*, Computational Linguistics, Volume 19, Number 2, pp. 331-358, 1993.

[Roget72] Roget P.M., *Roget's Thesaurus of Synonyms and Antonyms*, University Books, London, 1972.

[ShangWu84] Shang Wu 務印書館, 《現代漢語詞典》, 中國社會科學語言研究所詞典編輯室編,高務印書館, 北京, 1984.

[Tang94] Tang H.Y., Yao T.S., Kou Y.X., *The Lexical Semantic Driving Algorithm in Language Processing*, Proceedings of the ICCPOL' 94, Chinese Language Computer Society, pp. 333-338, 1994.

[Tong93] Tong X., Huang C.N., and Guo C.M., *Example-Based Sense Tagging of Running Chinese Text* , Proceedings of the Workshop on Very Large Corpora, June 22, 1993, Ohio State University, pp. 102-112, 1993.

[Tsinghua92] Tsinghua University 清華大學, 《漢語詞性自動標注系統》, 清華大學計算機科學與技術系, 1992.

[Tsutsumi93] Tsutsumi T., *Word-Sense Disambiguation by Examples*, Natural Language Processing: The PLNLP Approach, Kluwer Academic Publishers pp. 263-272, 1993.

[Vossen89] Vossen P., Meijs W., and Broeder M., *Computational Lexicography for Natural Language Processing: Meaning and Structure in Dictionary Definitions*, Longman, pp. 171-192, 1989.

[Wang90] Wang Y.C. 王永成, 《中文信息處理技術及其基礎》,上海交通大學, 1990.

[Weischedel93] Weischedel R. et al., *Coping with Ambiguity and Unknown Words through Probabilistic Models*, Computational Linguistics, Volume 19, Number 2, pp. 359-382, 1993.

[Wilks89] Wilks Y. et al., *A tractable machine dictionary as a resource for computational semantics*, Computational Lexicography for Natural Language Processing, Longman, pp. 193-228, 1989.

[Wilks93] Wilks Y. et al., *Providing Machine Tractable Dictionary Tools*, Semantics and the Lexicon, Kluwer Academic Publishers, pp. 341-401, 1993.

[Wong95] Wong K.F. et al., *A Parallel Approach for Identifying Word Boundaries in Chinese Text*, Department of Systems Engineering and Engineering Manamgent, Chinese University of Hong Kong, 1995. (submitted for publication)

[Yu93] Yu S.W. 俞士汶, 《計算語言學一教學參考資料》,北京大學計算機科學技術系-北京大學計算語言學研究所, 1993

[Yuan94] Yuan C.F. et al. (1994). *Combining Rules and Frames in Chinese Natural Language Parsing*, Technical Report, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong , 1994.

[Yuan93] Yuan C.F. and Huang C.N 苑春發及黃昌寧, 《關於語料庫和句法知識自動獲取的研究》, 清華大學學報, Vol. 33, pp. 71-76, 1993.

[Zernik91] Zernik U., *TRAIN1 vs. TRAIN2: Tagging Word Senses in Corpus*, Proceedings of RIAO' 91, pp. 567-585, 1991.

[Zhang92] Zhang P. 張普, 《漢語信息處理研究》,北京語言學院出版社, 1992.